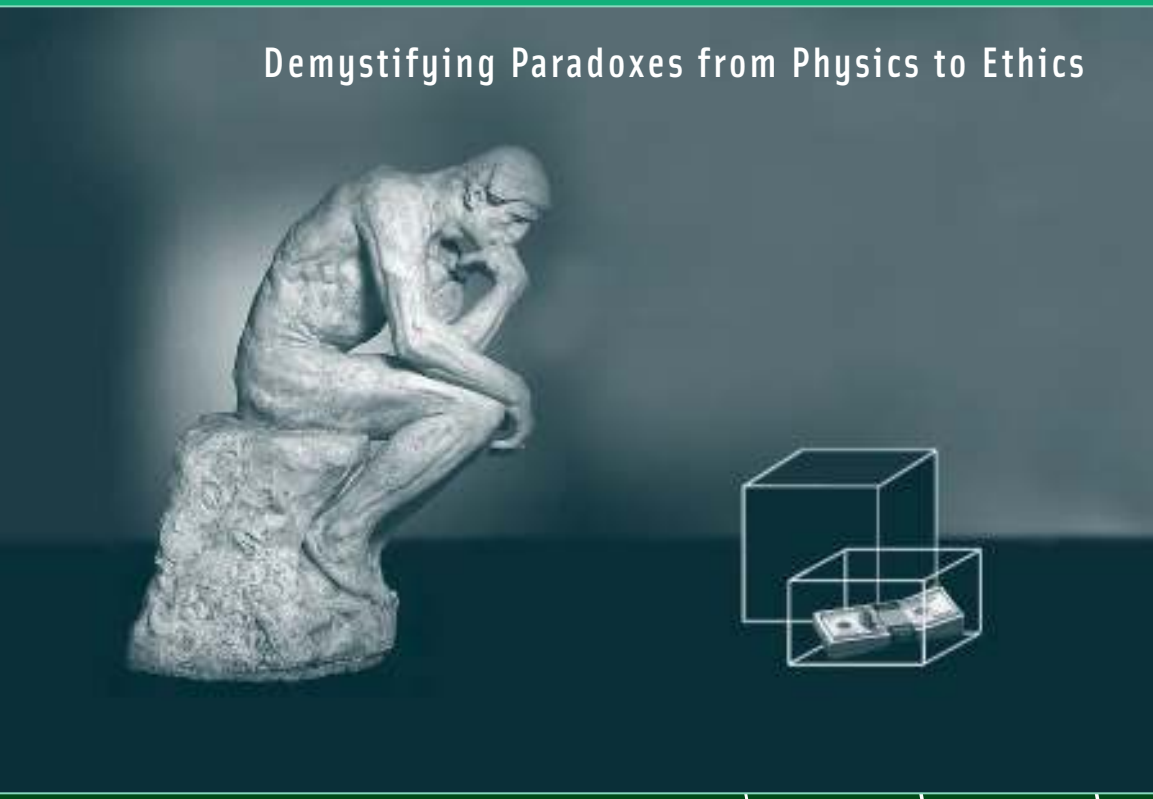
Gary L. Drescher

Good and Real

Demystifying Paradoxes from Physics to Ethics



Good and Real

Good and Real

Demystifying Paradoxes from Physics to Ethics

Gary L. Drescher

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Stone Serif and Stone Sans on 3B2 by Asco Typesetters, Hong Kong, and was printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Drescher, Gary L.
Good and real : demystifying paradoxes from physics to ethics / Gary L. Drescher.
p. cm.
''A Bradford book.''
Includes bibliographical references and index.
ISBN 0-262-04233-9 (hc : alk. paper)
1. Mechanism (Philosophy). 2. Philosophy of mind. 3. Ethics. I. Title.
BD553.D74 2006 146′.6—dc22 2005056169

10 9 8 7 6 5 4 3 2 1

In memory of Irving E. Drescher

Yet the magical imagery born of the mind
is just the same carved out of stone.

—Thomas Zimmerman

Contents

Preface

I came of age in the provocative environment of the MIT Artificial Intelligence Laboratory in the 1970s, a time of excitement and optimism about the imminent prospect of understanding and replicating the machinery of the human mind. To put it mildly, this quest has turned out to be harder than then expected; in retrospect, trying to construct human-scale intelligence using a one-megabyte, one-megahertz mainframe computer seems almost whimsical. Nonetheless, the continuing important effort has already stimulated myriad advances in science, engineering, and philosophy. It is the last of these domains that I turn to here.

Thinking about building minds inspires diverse, challenging questions about mind and reality. This book seeks to integrate several lines of inquiry that attempt to reconcile the mechanical nature of the physical universe with aspects of human nature involving consciousness, choice, time, and ethical right and wrong. Key aspects of our own nature certainly don't *feel* purely mechanical and material; I try to explain how our feelings may mislead us in that regard. As my thoughts on these matters receive scrutiny, I hope to encounter eventual confirmation or refutation of some of the claims staked out here—admittedly with more enthusiasm for the former possibility, but either way would constitute progress.

During the protracted incubation of these ideas, I've benefited from the advice and (often) encouragement of many individuals. Indeed, a number of friends and colleagues, having generously trudged through several evolving versions of this work, have shown greater patience for the subject matter than I myself have been able to summon at times. Of course, I cannot necessarily claim the endorsement of those who have lent a hand—many of whom, in fact, express disagreements ranging from the subtle to the apoplectic.

Recently, Daniel Dennett—whose influence permeates this book—was kind enough to invite me to be a visiting fellow at the Center for Cognitive Studies at Tufts University. There, I was fortunate to have valuable discussions with him and with Gabriel Love, Will Lowe, Oliver Selfridge, Rodrigo Vanegas, and others. I am also indebted to Uri Wilensky for his ideas, suggestions, and support; and likewise to Jim Davis (James R. Davis, Ph.D., MIT 1989, to distinguish him from the homonymous multitudes), who provided detailed, chapter-by-chapter comments on a draft of this book.

Part of the book's core technical content, chapter 4, is a revision of my paper "Demystifying Quantum Mechanics: A Simple Universe with Quantum Uncertainty," previously published in *Complex Systems*. Another core portion, chapters 5 through 7, addressing the ramifications of Newcomb's Problem, was circulated separately in draft form. The following people, in addition to those just mentioned, have offered significant comments on one or both parts of this technical core, or on drafts of the entire book: Phil Agre, Richard Amster, Jonathan Amsterdam, Manor Askenazi, John Batali, Alan Bawden, Max Behensky, Ken Binmore, Mario Bourgoin, Pete Cann, David Chapman, Tom Clark (of the Center for Naturalism), Eric Cohen, Judy Feldmann (of the MIT Press), Ed Hardebeck, Danny Hillis, Ian Horswill, Tom Knight, Joachim Krueger, Stan Kugell, Margaret Minsky, Marvin Minsky, Ron Rivest, Deb Roy, Jerry Roylance, Bill Silver, Brian Silverman (who first called my attention to Newcomb's Problem), Erric Solomon, Richard Stallman, Oliver Steele, Tom Stone (of the MIT Press), Gerry Sussman, Pablo Tamayo, Christopher Taylor, Lucia Vaina, Dan Weld, Joe Yarmus, Ramin Zabih, several anonymous referees, and no doubt others, whom I regret neglecting.

Stripped of any intellectual pretension, philosophy just grapples with the ancient, heartfelt questions *What the fuck is going on here?* and *What the fuck am I doing here?*—what is the nature of reality, and what is the nature and purpose (if any) of our place in it? We all strive to answer such questions somehow or other. Here goes one more try.

Gary Drescher
April 2005

Good and Real

1 Introduction: Framing the Big Picture

1.1 Us and the Universe

Philosophers sometimes seem obsessed with arcane puzzles whose solutions, even if available, leave us no wiser about things that matter. Some puzzles, on the other hand, can enlighten us, because they expose and explore confusions that are central to our grasp of reality. By untangling the principles showcased by such riddles, or even just by trying to, we can learn much about ourselves and our universe. This book presents, and tries to resolve, some illuminating paradoxes that arise when we try to reconcile our subjective impressions of our own existence—and the universe's—with what we know scientifically about ourselves and the world.

I want to discuss a particular view of reality: that the universe is a machine, its behavior specifiable by a simple set of regularities (the details of which are not quite known yet, although physicists may be getting close). I believe this view is correct, but it is its coherency, not its truth, that I try to defend here; its truth is an empirical matter for which physical scientists have responsibility. What I wish to address is: *given* that the recent millennia of empirical data have pointed increasingly toward a mechanical universe (with the notable apparent exception of quantum physics, addressed in depth later in this book), how can the mechanical viewpoint be reconciled with several apparent contradictions, especially about matters that most concern us—matters of choice, ethics, consciousness, and other aspects of our own nature? Were such a reconciliation at hand, I would expect the burgeoning mountain of empirical evidence to prevail easily, establishing the mechanical model beyond significant doubt.

Laplace famously proposed that the universe is a clockwork-like mechanism. Simple patterns of events, so-called *physical laws*, describe how

objects move and interact. Those laws, together with the complete, detailed state of the universe at a given moment, specify the state of the universe at all times by saying exactly how the universe changes from one moment to the next. More recently (and more speculatively), Edward Fredkin (1990), Stephen Wolfram (2002), and others have proposed that the universe's clockwork may be a *cellular automaton*: a discrete grid (like a chessboard), with simple, uniform, deterministic, rules that say what happens in the next state, as a function of the current state. In a cellular automaton, all states are discrete, digital, in contrast with the smooth, analog states of the Laplacian model. But that difference is irrelevant to most of what I discuss here, so I help myself to convenient illustrations of both types.

A clockwork universe is at odds with a number of powerful intuitions. The idea of a mechanical universe (together with some of the known details of this universe's particular mechanism) presents a series of problems that this book examines:

- *The problem of consciousness* We are *aware* and we *feel* and we *care*. It is not evident how any arrangement of inanimate mechanical components can do so.
- *The problem of ethics* We perceive a difference between *right* and *wrong*. But it is not evident where such a difference could come from. How can a purely mechanical system give rise to a way something *should* be, in contrast with how it simply *is*?
- *The problem of choice* If the entire future of the universe is already determined by the past (or even—as some interpretations of quantum mechanics suggest—by the past plus some random coin tosses), then are all choices futile? Is it just an illusion that our choices have any sort of efficacy?
- *The problematic flow of time* The universe always seems to have a *present* state, a *now*, that seems to move forward in time. But known physical laws describe no such motion. Rather, the entire past, present, and future are sitting statically in spacetime, like the collection of frames that constitute an already-completed reel of film. Why, then, does a flow of time seem very real to us?
- *The problematic asymmetry of time* The particular physical laws that are found to hold throughout our universe are *time symmetric*. Yet somehow, the universe that accords with those laws is highly asymmetric in time: we remember the past, but not the future; we cause things in the future, but

not the past; entropy increases in the future; and (as just noted) time itself gives the impression of flowing forward, but not backward. How can symmetric laws produce asymmetry?

- *The problem of existence* Why does the universe have its particular mechanical laws? Why is there this universe, rather than some other or none?

- *The problem of quantum-mechanical uncertainty* According to quantum mechanics, any unobserved particle has no single definite state, but rather is in a *superposition* of possible states. Yet an observation of the particle always shows a single definite state, seemingly selected at random from the superposed possibilities. But how can a mechanical particle play hide-and-seek by pretending to have a single state whenever we observe it, even though it gives evidence of having many states when it is *not* observed?

This book may have far more scope than it deserves. It presumes to address some of the most profound questions ever framed—questions about the nature and value of existence, both physical existence in general and human existence in particular. By training, I am a computer scientist, not a philosopher or a physicist. Yet I delve here into questions in all those fields, especially at their intersection.

Tackling one or another of the subjects examined here is respectable. But assembling so many of them may seem grandiose, except perhaps for my sincere acknowledgement of the tentativeness of the endeavor. In any event, the latitude is not capricious—matters of fundamental importance, whether about physics or about human consciousness, have ramifications that intertwine (or at least appear to), making it difficult to adequately address each such matter in isolation. And I am hopeful that at least some of the ideas I present along the way are novel, interesting, and even approximately correct.

The disparate topics of this book have a unifying theme: that questions about reality, to the extent that they are meaningful, have objective, rational answers. This theme is recognizably outdated, a remnant of the classical physics and logical-positivist philosophy of a century or more ago. I do, in fact, believe that the positivists—whose position was defined by their commitment to empirical evidence and logical argument as the means for finding truth—were in many ways on the right track, but lacked some crucial technical tools to make their position work:

- Without the insights that the fields of cognitive science and artificial intelligence later developed, positivists could not account adequately for the objective existence of subjective, conscious experiences. Chapter 2 below reviews some of these considerations.
- Without Everett's relative-state interpretation of quantum mechanics (chap. 4), the positivist construal of physics fell into disarray.
- The positivists also lacked a viable account of what are known as *subjunctive* or *counterfactual* propositions (chaps. 5 and 6), which speak of hypothetical alternatives to the way things really are. Contemplating such alternatives is crucial to thinking about *choices* and their consequences. (A subjunctive statement is of the form *If X were the case, then Y would be the case*.)
- And in a surprising twist, subjunctive reasoning may turn out to provide a basis for ethical truth (as outlined here in chap. 7), a domain where positivism could find little purchase.

Without the necessary technical means, the positivists' efforts could not enjoy a coherent foundation—especially with regard to ethics—and the despair of decades of foundering helped give rise to postmodern philosophy, which (broadly speaking) rejects the quest for objective truth and construes reality as a social construction, a creation of culture.

As do many others, I regard the postmodern surrender as unwarranted. In reality, the earth is round, $E = mc^2$, $e^{i\pi} = -1$, and the capital of France is Paris. All these facts are true regardless of what, if anything, any cultures hold about them—except, of course, the last, to the extent that it is a statement *about* a cultural matter. Yes, the very concepts of earth, energy, and so forth, are social constructs, in the sense that the ideas themselves (as opposed to their referents) were developed by societies. They are also cognitive constructs, in the sense that they were developed by individuals' minds; biological constructs, in the sense that evolution eventually built the brains that give form to those concepts; cosmological constructs, in that our brains, like all other things in the universe, are part of the continuing unfolding of the big bang; and so forth, with different "-ologies" relevant from different legitimate points of view.

And yes, few of us would know that the earth is round, and so forth, if not for being instructed so by our cultures. Furthermore, the only reason we may *care* about these facts is that they pertain to our human needs or

wants, many of which are enmeshed in our cultures. Nonetheless, there is a fact of each of these matters, and if a culture can know and teach these facts, it is only because individuals in that culture have engaged in the processes of rational inquiry—for example, performing experiments, proving theorems, proposing analogies, elaborating narratives—that can discover and convey the truth.

Like many, I believe the two most important achievements of human intellect have been the realization of the essential character of the physical universe, and the appreciation of the central principle of ethics—understanding what is, and what should be:

- Arguably, the essential character of physics is that our universe is mechanical, with a small number of fundamental kinds of building blocks that behave in total accordance with a small number of concisely, formally expressible patterns. The mechanical nature of reality was first hinted at by the conspicuously quantifiable regularity of the stars and planets. Some ancient Greeks suspected, and the last few centuries of science have confirmed, that astronomy is paradigmatic, its elegant regularity literally universal.
- Arguably, the essential character of ethics is approximated by the Golden Rule: you should treat others as you'd want others to treat you. This principle, too, has ancient origins, but demonstrating its validity has proved more elusive, despite modern methods of inquiry.

At the dawn of the last century, physics took a wildly unexpected turn with the discovery of quantum phenomena. These phenomena are bizarre in ways that far exceed their mere apparent nondeterminism; quantum phenomena seem to challenge the very notion that there is an observable, objective, mechanical universe. Instead, quantum experiments seem to show that the very act of observation—by some accounts, specifically human or conscious observation—is what gives rise to the states that are being observed, which cannot correctly be said to have existed prior to or apart from the act of observation. Ironically, the positivists' commitment to grounding knowledge in observations—a grounding that was supposed to be a recipe for objectivity—dovetailed here with a relativist perspective whereby the observer effectively constructs all reality instead of just discovering it. (This convergence between positivism and relativism is a bit

reminiscent of the way extreme left-wing politics sometimes resembles its right-wing counterpart.)

An observer-dependent interpretation of quantum phenomena does not compel a postmodern rejection of objective truth per se. Conversely, postmodernism certainly does not compel any particular theory of physics. Still, at least historically and allegorically, the renunciation of the paradigm of objectivity by quantum physicists has heralded and supported the more general postmodern movement,[1] including a relativist approach to matters of right and wrong. After all, if physics—the very poster child for the paradigm of an objective and rationally knowable state of affairs—has had cause to abandon that paradigm, then it is scarcely likely that more human-oriented matters such as ethics, which hadn't looked very convincingly objective in the first place, would turn out to be so. Popular writings often draw a connection between quantum mechanics and the supposed subjectivity of reality, sometimes in concert with especially dubious claims about the putatively fundamental influence of human consciousness on low-level physical phenomena, and vice versa.[2]

But in 1957, the physicist Hugh Everett proposed a reinterpretation of quantum phenomena that fully reconciled those phenomena with the old mechanical (and even deterministic) paradigm. Everett's so-called *relative-state* formulation has been gaining acceptance among physicists, impeded, perhaps, by the difficulty of translating between its complex mathematical model and the English glosses that help connect to one's intuitions.

In chapter 4 of this book, I present a simplification of Everett's model that serves to demonstrate his point with formal precision, but using a

1. In a famous spoof, the physicist Alan Sokal submitted to the postmodern cultural-studies journal *Social Text* an article (1996) that invokes quantum gravity to help dismiss the "dogma … that there exists an external world, whose properties are independent of any individual human being and indeed of humanity as a whole." In order to test what he suspected was a laxness of intellectual standards, Sokal intentionally crafted a grossly nonsensical paper (for example, he proposed that the mathematical axiom of equality and axiom of choice bear on social issues regarding gender equality and the "pro-choice" stance—a correspondence supported by nothing but the coincidental similarity of the corresponding names). Confirming Sokal's suspicion, the editors of *Social Text* found the article sufficiently sound that they published it.

2. Fritjof Kapra's *The Tao of Physics* (1975) is a classic example. *Quantum Consciousness* (Wolinsky 1993) and *The Self-Aware Universe: How Consciousness Creates the Material World* (Goswami et al. 1995) are others.

stripped-down formalism that requires no advanced math or physics. The original interpretation of quantum mechanics seemed to force a rejection of the idea of objective physical reality, which rejection then supported, at least indirectly, the renunciation of objectivity in other spheres as well. Conversely, I hope that Everett's insight, if it can be made accessible to nonphysicists, can serve to illustrate how a missing technical linchpin can make the enterprise of seeking objectivity seem so hopeless as to be abandoned as impossible—until the elusive technical fix comes along.

If the postmodern retreat from objectivity is misguided with regard to physics, it is disastrous with regard to ethics. True, most of us inherit most of our ethical convictions from our cultures, as we do our scientific beliefs. But our inherited ethical convictions had better be able to be grounded in a fact of the matter, just as in physics. For if not, there is no reason for individuals, or cultures, to keep from adopting arbitrarily selfish or cruel standards, to the extent that such standards benefit those who adopt them.

To regard culture as the ultimate arbiter of ethics is insidious. Insofar as a culture is largely benevolent, moral authority supposedly grounded in that culture may look plausible. Yet if culture per se is thought to define what is right, then a drift toward ''manifest destiny'' is hard to resist. After all, we can rather reliably predict that powerful cultures will conquer. If their doing so will make it right, then that is what is right.

The danger of abandoning attempted objectivity is of course historical, not hypothetical. It is instructive here to contrast the liberal humanism of Bertrand Russell with the racist fascism of Martin Heidegger. Arguably, both of their political stances were deeply and thoughtfully rooted in their respective prodigious philosophies.[3] But Russell's positivism committed him to an objective reality, whereas Heidegger laid much of the foundation for recasting supposedly objective fact as mere contrivance, reflecting the practices and purposes of the contrivers rather than being tied to the way things really are. Such a view makes the world—including its ethical facets—seem arbitrarily malleable. When a philosophy thus relinquishes its anchor in reality, it risks drifting arbitrarily far from sanity.

3. Concerning his own philosophy and politics, Russell expounds on the connection in the introduction to his *History of Western Philosophy* (1945), with elaborations throughout the book. Regarding Heidegger's philosophy and politics, see, e.g., Fritsche 1999.

Steven Pinker, in *The Blank Slate* (2002), observes that many students, steeped in postmodern ethical relativism, hedge even their most basic moral judgments,[4] saying things like "Our culture places great value on treating others well," but stopping short of the idea that harming others could be wrong even if everyone in our culture thought it was okay. Similarly, I recall a political discussion years ago in which a fellow graduate student carefully avoided claiming that Nazism was wrong, saying instead how glad he was that Nazism had "fallen out of fashion," and how he hoped it never becomes fashionable again. Here was a smart, decent person who could find nothing more decisive to say against Nazism than could also be said about polyester leisure suits! And he saw nothing odd or disturbing about that parity. (To be fair, I'm sure he felt much more passionately about the Nazis, but still.) How do we educate children to be humane if the most stirring reason we can muster is a wish that their tastes won't go retro?

Ethical relativism is an understandable reaction to centuries of the opposite excess, in which the powerful have proclaimed that their self-serving values—entitlement to conquer vulnerable peoples, and to the subservience of slaves, women, and the poor, for instance—are moral absolutes that others are obliged to yield to and could rightly be forced to yield to. Ironically, these putatively absolute moral values have been rationalized by appeal to two quintessentially relativist sources: faith and tradition. Faith, after all, is just a strong subjective feeling as to what's true. And tradition is just an appeal to ancient peer pressure (typically intimately interwoven with faith). Deriving moral foundations from tradition leads us to conform to the beliefs and behavior of people who—though they did their best with the knowledge then available—lived long before even a high-school education existed. But as every child learns (or should), the claim that "everyone else does it," or has always done it, is a poor justification for one's conduct.

Thus, the faith-and-tradition position combines the worst of both worlds: it rests the certitude of absolutism on a flimsy relativist foundation, often with catastrophic results. To its credit, postmodernism rejects the cryptorelativism of absolutist faith- and tradition-based ethics (as previously did

4. The terms *moral* and *ethical* sometimes have different connotations, but throughout this book I use them interchangeably to refer simply to matters of right and wrong.

Nietzsche[5] and others), but without being able to replace it with an objective alternative.

Unfortunately, attempts to construct objectively grounded theories of ethics tend to end up with theories of selfishness instead. A constellation of *self-interest* theories posit, in one way or another, that individuals should just act in pursuit of their own personal interests, and that an acceptable approximation to what we think of as responsible, ethical behavior will follow. Such theories include ethical egoism, sociobiologically based ethics (ethics via reciprocal altruism), and libertarian capitalist ideology (an extreme example of which is the so-called Objectivist ethical theory of Ayn Rand [1964]).

The intellectual motivation for self-interest theories is readily comprehensible (as too are less laudable motivations). To the extent that treating others well tends to cause oneself to be treated well, one has a solid rational basis for treating others well; and other solid rational bases for doing so are distressingly hard to come by. It is easy to assert an obligation to refrain from harming others, and (except for sociopaths) easy to feel emotionally compelled by such an obligation, but it is far harder to say what, exactly, would be mistaken about totally denying the obligation—unless one has already presumed some undemonstrated ethical principle that is more or less tantamount to the desired conclusion.

Not surprisingly, however, although self-interest often motivates respectful conduct, it sometimes motivates predatory behavior instead. When self-interest diverges from respect for others, self-interest theories variously declare the unkind behavior acceptable, or else implausibly deny the divergence, offering contorted reasons that the unkind behavior must supposedly cause prohibitive harm to one's own interests. Perhaps in desperation to provide *some* objective foundation for ethics (or perhaps just to rationalize selfishness), self-interest theories cling to an inadequate technical basis for ethics, resulting in theories that are objective but false.

Given the apparent dichotomy between objectively justified selfishness and nonobjective altruism, it is no wonder—and indeed, in some ways

5. "... since Plato, all theologians and philosophers have followed the same path—which means that in matters of morality, instinct (or as Christians call it, 'Faith,' or as I call it, 'the herd') has hitherto triumphed.... I hope to be forgiven for discovering that all moral philosophy hitherto has been tedious and has belonged to the soporific appliances" (Nietzsche 1917).

fortunate—that the notion of objectively demonstrable truth about ethics has yielded. Still, as just noted, relinquishing objectivity brings stark problems of its own. Ultimately, neither branch of this false dichotomy is correct or tolerable, and the existing precarious compromise—emphasizing sometimes one, sometimes the other, with no principled coordination—is better than either alone, but is no substitute for a true alternative.

In the domain of ethics, as in quantum physics, I believe there is a technical fix that provides a foundation for ascertaining a fact of the matter, for arriving at rationally derived, objectively true answers to questions in that domain. For ethics, what turns out to be needed, I argue, is a subtle technical development in the theory of subjunctive or counterfactual reasoning, which I address in chapter 5 of this book. Very briefly, the connection between subjunctive reasoning and ethical theory arises (I argue in chap. 7) because if you were to have reason to decide to behave benevolently to others (even if you could profit from behaving otherwise), then others in symmetric situations would have reason to decide to behave benevolently toward you—even if the decisions are made independently (such that you and the symmetrically situated others have no knowledge of one another's decisions). I argue that this subjunctive relation (regarding what would be the case if you were to behave a certain way) rationally motivates you to act toward others as you want others to act toward you—thus vindicating the familiar golden-rule or categorical-imperative intuition—even when your behavior cannot *cause* any reciprocal benevolence.

The resulting ethical theory (elaborated in chap. 7) is still, in some sense, grounded in self-interest, but in a way that can justify acting for others' benefit even when doing so causes nothing but harm to one's own interests. The theory can even justify acting to uphold some principle in a case where doing so causes nothing but harm to *everyone's* interests, contrary to what utilitarianism advocates.

Of course, the specific technical fix proposed here for ethics is entirely distinct from the fix for physics. What they have in common is just that both have aspects that, at face value, are intensely counterintuitive; yet both ultimately reaffirm common sense in a way that the nonobjective alternatives cannot. The two disparate subjects also intersect in that, as mentioned above, some authors have invoked nonobjective interpretations of quantum mechanics to support nonmechanical theories of the mind and its value. But I argue in chapter 4 that the unnoticed intrusion of mis-

taken theories of mind into the realm of physics theory is what led to non-objective interpretations of quantum mechanics in the first place!

To view the universe and its contents—including us—as machines strikes many people as implausibly and unpleasantly cold, a peculiar denial of our true nature—a nature in which emotion figures prominently in who we are and how we come to know the world. And indeed, my intention is not at all to deny the existence or importance of emotion, of feeling and passion, but rather to help promote the view that—contrary to what many believe—those phenomena are *compatible* with a mechanistic view. That is, there are some purely mechanical entities—including you and me—that are capable of reason and emotion. Seeing how it is possible for some machines to have such capabilities is important to our self-understanding. And distinguishing what is true from what merely *feels* true is important to our understanding of *anything*.

Our culture's pervasive skepticism about reason and mechanism is amply proclaimed in our popular entertainment. In *Star Wars*, a mentor instructs his blindfolded student to *trust his feelings*, his intuition, as a substitute for the missing sensory information. In the film's mystical fantasy world, that advice turns out to be sound, which makes for fun storytelling.

But reality is quite different. In the early days of aviation, for example, pilots flying inside clouds would regularly lose control of their aircraft and crash. Unable to see the ground or the sky, the pilots literally could not tell which way was up. They relied on their sense of balance and their overall spatial intuition. But as an airplane banks, its flight path curves, and centrifugal effects keep the apparent downward direction pointing straight to the floor of the airplane. To its occupant, the cloud-enshrouded airplane still *feels* level even as it banks and dives more steeply.

Today, safe flight inside clouds is possible using gyroscopic instruments that report the airplane's orientation without being misled by centrifugal effects. But the pilot's spatial intuition is still active, and often contradicts the instruments. Pilots are explicitly, emphatically trained to *trust the instruments* and *ignore intuition*—precisely the opposite of the *Star Wars* advice—and those who fail to do so often perish.

In fact, the pilot's spatial intuition is itself based on information from mechanical sensors in the pilot's body—sensors that provide visual, tactile, proprioceptive, and vestibular cues about spatial orientation. Ordinarily,

those sensors work well; without them, we'd be unable even to walk. But those particular sensors are inadequate when flying inside clouds—there, we need gyroscopes.

The plight of the pilot illustrates a crucial principle: rationally understanding how our feelings and intuitions are mechanically implemented can help us distinguish when our intuitions are trustworthy and useful and when, on the other hand, they mislead us—sometimes calamitously. The following chapters look into the underpinnings of some of our deepest intuitions—about consciousness, choice, right and wrong, the passage of time, and other matters—in an effort to draw a similar distinction.

1.2 Ground Rules and Terms of Discussion

Before the subsequent chapters delve into the ideas outlined above, there are a few procedural matters to discuss—matters that concern the very process of thinking about and debating the issues at hand. These matters include the objectivity and absolutism (or not) of truth, the role of definitions in reasoning, and the nature of paradoxes and their resolutions. Although these concerns are elementary, attendant confusion can easily derail informal discussions (or even rigorous analyses) of philosophical topics. Hence, here is an attempt to address these concerns explicitly before getting underway.

1.2.1 Unbending the Truth

Is truth objective or subjective? And (a distinct question) is it absolute or situation dependent? Often, the concept of absolute, objective truth seems misguided. Consider the statement *Chocolate tastes better than vanilla*. Surely that is just, as we say, a matter of taste—there is no fact of the matter as to which is better. Or (even apart from subjective judgments) consider the statement *I am more than one meter tall*. Clearly that statement can be either true or false, depending on who says it and when—its truth is dependent on the circumstances of its utterance.

On closer inspection, though, truth turns out to be more rigid than these considerations suggest. Seeming variability comes about in part when a statement is vague enough to take on any of a range of meanings; the statement can then be true with respect to some of its meanings and false with

respect to others. But if the meaning is pinned down sufficiently, that apparent variability vanishes. *Chocolate tastes better than vanilla* is unambiguously true or false if it refers to a particular person's net preference at a particular moment.

Thus, the preference for chocolate or vanilla is indeed a subjective matter, but that doesn't make truth itself subjective. Rather, the *subject matter* of this particular true (or false) statement is subjective. But the statement itself (if pinned down by saying whom and when it refers to) is a factually true (or false) claim *about* that subjective subject matter.

Similar considerations apply to the absolutism versus situation-dependency of *I am more than one meter tall*. Yes, its truth depends on who says it and when. But who says it, and when, is part of the very meaning of the statement—in effect, it's asserting something different when uttered under different circumstances. Among those different assertions, we find a mixture of truth and falsehood. But each of those assertions individually is simply true or false.

Indeed, the question of whether truth is absolute or situation dependent is incoherent, because any truth can be expressed in either absolute or situation-dependent terms. By building enough conditions into the statement itself, you can make it absolute—for example, *An unsupported object near the surface of the earth accelerates downward at about ten meters per second squared*. Or by omitting relevant conditions, you can make the statement depend on the circumstances of its utterance—*An object accelerates at about ten meters per second squared*.

Logicians refer to sufficiently pinned-down statements as *propositions*, each of which is definitively true or false (even if we have difficulty ascertaining which). An ordinary English *statement* may correspond to a variety of propositions, some of which may be true and others false. In particular, statements that depend on the time, place, or identity of the speaker—such as *I am more than one meter tall* or *The object I'm pointing to is a chair*—are called *indexical* statements. An indexical statement is, in essence, a function whose input is a time and place and whose output is a proposition that refers specifically to that time and place.

The question of subjectivity and situation-dependency often comes up in discussions of ethical matters. The whole question of ethical situation-dependency is, I believe, a red herring, as in the nonethics examples above.

- Any proposed ethical principle that can be expressed in absolute terms—for example, *It is always wrong to kill*—can be expressed instead in situation-dependent terms—for example, *It is wrong to pull the trigger in a situation where the gun is loaded and pointed at someone's head.*
- The converse holds as well. If nothing else, any situation-dependent principle could be transformed into one that is always true by stating a new principle that exhaustively lists the situations in which the dependent principle is true, and asserts that the dependent principle is true under precisely those circumstances. (That is, if *X* is true if and only if *Y* is true, then the new proposition *If Y, then X* is always true.)

Insofar as any ethical principle can be stated in either absolute or situation-dependent terms, saying that ethical truth is absolute (or that it is situation dependent) is like saying that ethical truth is English or Chinese, depending on which language it's expressed in. But the two are intertranslatable, so any truth can be expressed either way. Thus, neither being English versus being Chinese—nor being absolute versus being situation-dependent—is a property of the truth itself. It is merely a property of the way we choose to express the truth.

Ethical subjectivity is a more difficult matter. Ethical *relativists* consider all ethical truths to be subjective—whether it is morally right to do *X* depends on whether an individual (or a culture, depending on the brand of relativism) *believes* that *X* is morally right. Taken literally, though, that formulation is vacuously circular. If what is morally right is nothing other than what you believe is morally right, then the second occurrence of "morally right" (i.e., the object of your belief) *also* just refers to what you believe is morally right—so then, what is morally right is what you believe you believe is morally right, which is what you believe you believe you believe is morally right ... But that substitution keeps on going forever, and we never get to what it is that you supposedly believe about *X*.

To avoid this infinite regress, to make relativism coherent (but still not necessarily correct), *X* being morally right has to correspond not to believing that *X* is right, but rather to some other property—believing something else about *X*, or having some attitude about *X*. For example, it might be proposed that *X* is morally right if, on the whole, you desire to do *X* and feel comfortable about it, or if your culture promotes your doing *X*, or whatever. Such proposals are at least meaningful.

But since it is possible for individuals or groups to become comfortable with and promote almost any self-serving behavior, relativist proposals—relegating ethical truths to the realm of subjective individual or cultural preferences—tend to collapse into ethical nihilism—the view that nothing is wrong, that anything you want to do (individually or as a culture, depending on the flavor of relativism) is thereby ethically permissible.

Of course, such a position is no longer a subjective claim about ethics. Ethical nihilism is the doctrine that any action you choose to take is as right as any other—a perfectly objective claim. But chapter 7 below argues that it turns out to be an objectively *false* claim, and that a rational foundation exists for a nonnihilist, nonrelativist ethics.

1.2.2 Definitions and Semantic Sleight of Hand

Apart from apparent subjectivity and situation-dependency, our latitude about how to define words presents yet another challenge to the notion of objective, absolute truth. Consider a statement such as *Most frubles are green*—is that true? Evidently, it depends on how we define the made-up word *fruble*. If *fruble* means *leaf*, the statement is indeed true. But if *fruble* means *lump of coal*, the statement is false.

But the uncertainty about the truth of *Most frubles are green* is cleared up once we specify what *frubles* are. The definition of *fruble* (or of any word) is arbitrary, of course—a definition is just a decision to represent a particular concept by a particular series of syllables or signs or squiggles. The definition can be anything we choose. But the arbitrariness of definitions doesn't make truth arbitrary. Rather, it just means that in order to understand which proposition it is whose truth we're being asked about, we need to know what the words mean. Once again, it is just a matter of pinning down the meaning in order to pin down the truth.

Even though definitions are arbitrary, some may be more confusing than others, especially if you choose to define a common word differently than other people usually do. In that event, it behooves you to point out your idiosyncrasy, lest you create confusion. It's fine, for instance, if you want to use the word *automobile* to refer to a shoehorn rather than to a motorized conveyance. But you'd better explain your unconventional usage when you ask someone to put your automobile in the hall closet.

As obvious as the need for consistency of definitions seems when laid bare, it often seems to be a source of profound confusion. What can happen is that a word will be used sometimes with one meaning, sometimes with another, with the transition unnoticed. Consider a preposterous example: someone who believes that leaves have the same color as coal can appear to support that belief—without the need for any physical evidence—simply by establishing the truth of *Most frubles are green* using the *leaves* meaning, then switching to the *coal* meaning and relying on the statement's previously established truth.

That semantic sleight of hand is unlikely to be convincing in this fanciful case, because it is too obvious that whether leaves are the same color as coal is a substantive, empirical matter, not something that could possibly be decided at will by redefining a word in such a way as to conflate two meanings. But the same sleight-of-hand confusion arises surprisingly easily when a familiar word carries a connotation that smuggles in an unspoken meaning, which then gets conflated with some explicit definition offered for the word.

Consider, for example, the definition of *life*. Many older textbooks defined a living entity as something that exhibits a handful of telltale abilities: respiration, assimilation, reproduction, and so on. Occasionally, a novel entity (some complex molecule, or a computer program or a robot) boasting a subset of the telltale signs will create debate among laypersons—though seldom among modern biologists—about whether it is a form of life. The debate does not focus on whether the entity meets a given definition of *life* (which is usually fairly clear), but rather whether the proposed definition is ''correct,'' or whether it needs to be revised in light of the new entity.

Insofar as definitions are arbitrary, there is nothing substantive here to debate. If there seems to be some substantive consequence to a decision to define a word one way or another, that is a sure sign that the word is already smuggling in an implicit, intuitive definition. Then the real underlying question is whether the implicit definition does or does not coincide with the proposed explicit definition—which is indeed a substantive question.

In the case of *life*, for example, a vitalist—who believes that some special force animates living things, rather than their just being arrangements of ordinary, inanimate particles—may implicitly define *life* as the possession

of that special vital force. With that meaning silently smuggled in, the seeming question of the correct definition of life is really the question: is a given explicit criterion sufficient (or necessary) to indicate the presence of the special animating force? That's a perfectly substantive question, and one whose answer does not depend on arbitrary decrees of definitions. But biologists, who have long since abandoned vitalism, see nothing interesting to debate there, since the special animating force turns out not to exist at all.

There need not be any deliberate deception or dishonesty involved in the sleight-of-hand conflation of an unspoken, implicit definition with an explicitly proposed definition. On the contrary, in the absence of careful effort to avoid it, the mistake is easily made without even noticing the back-and-forth substitution. And nowhere does that happen more readily than in discussing matters of right and wrong.

Many definitions of *right* and *wrong* have been proposed. For example, *right* has been defined as *that which causes the greatest overall pleasure*, or as *that which respects the fundamental interests of all people (or all beings)*, or as *that which complies with scriptural claims about God's commands*, and so on. Or, as discussed above, *right* can be defined relativistically as *that which society promotes* or the like.

Different such definitions pick out different sets of choices as the morally right ones to make. Thus, much seems to hinge on the definition here. If, say, we assign the word *right* a utilitarian, greatest-overall-pleasure-causing definition (to pick one of the proposed definitions at random), then whether we should take a given action seems now to hinge on whether the action causes the greatest overall pleasure.

As mentioned above, whenever something substantive seems to depend on a choice of definition—for example, if whether to take a contemplated action seems to depend on whether the action falls within the scope of some proposed definition of *right*—we should suspect that a tacit definition is being smuggled in, and a sleight-of-hand substitution of the tacit definition for the explicit one is occurring. Here's a good diagnostic technique: define some made-up word in place of the familiar one that is being defined, and see what apparent difference that substitution makes.

In the case of a proposed definition of *right*, the difference is striking. Again, consider the utilitarian definition for the sake of illustration, even if (like me) you do not believe in purely utilitarian ethics (remember, a

definition is just an arbitrary association between a symbol and a concept; it has nothing to do with what is true or false about the world). If you're contemplating a given action, but are reminded that the action is not right, that objection seems germane. And in the case of the definition proposed here, that objection translates to the observation that the action does not achieve the greatest available overall pleasure.

Now suppose we keep the proposed definition of *right*, but we change the symbol that's being defined—instead of the word *right*, let's use the word *asdfil*. If you again contemplate taking the given action, but are informed that the action is not asdfil, your likely reaction would be: *So what? What does being asdfil have to do with whether or not I should take this action?* Unless you happen to be convinced of the correctness of utilitarian ethics, you have no reason to care about whether your actions have the arbitrary property defined as *asdfilness*.

Why, then, wasn't that our reaction when the word *right* was given the same definition as the word *asdfil*? If the two words are identically defined, why not be just as dismissive of what is right as we are of what is asdfil? The answer, I think, is that as we ordinarily use the word *right* (in the moral sense of right vs. wrong), we already implicitly define it to mean that which is not somehow mistaken for us to do (even if doing it fulfills any self-interested goals we may have). That is, if the action is right, then choosing the action does not somehow reflect some misunderstanding. When we propose an explicit definition as well for the word *right*, we then tend to equate the explicitly defined concept with the concept that is already implicitly attached to the word (roughly, the concept of nonmistakenness, in the above sense), without even noticing the sleight of hand.

Thus, we tend to think that whatever action meets the proposed explicit definition of right is not a mistaken action to take—there is no reason not to do it if we want to (and inversely, to do what is *not* right *is* somehow mistaken). But that equivocation is illusory. We cannot make two concepts equivalent (here, the explicit and implicit definitions of *right*) just by defining the same word to represent the two, any more than we can make leaves be the same color as coal just by defining the same word both as the color of leaves and as the color of coal.

Suppose, though, that we made a further effort to establish the equivalence of the implicit not-mistaken-to-do sense of *right* with, say, the utili-

tarian greatest-pleasure sense (once again selecting that particular sense at random for the sake of illustration):

- To pursue that effort, suppose we were to define *right* as that which is not mistaken to do *and* which brings about the greatest pleasure. Would that do the trick? It would not; given that conjunctive definition, we'd need to establish (somehow) that an act is not mistaken to do *and* that it promotes the greatest pleasure in order to show that it's right. Merely establishing the latter would not suffice.
- Well then, suppose instead we were to define *right* as that which is not mistaken to do *or* which promotes the greatest pleasure (using *or* inclusively, to denote *either or both*). In that case, we could indeed establish that an act is right merely by showing that the act promotes the greatest pleasure. But given that disjunctive definition, just establishing that something is right no longer establishes that it's not mistaken to do (it merely establishes that it is not mistaken *or* that it promotes the greatest pleasure).

The imperviousness of truth to these tricks of definition is reassuring. If concepts yielded to our attempts to equate them just by our proclaiming definitions in that manner, then definitions would be like magic spells, capable by their mere incantation of somehow rearranging the substantive facts of the world. Obviously, definitions have no such power. As long as we take care not to allow sleight-of-hand substitutions to go unnoticed, definitions are (at least in principle) inconsequential and need not be argued about.[6] In particular, for instance, defining *right* in the above utilitarian manner is inconsequential, if we're careful enough: if we fastidiously use just the stated definition, avoiding the smuggled-in implicit definition, then the stated definition tells us nothing about what behavior is or is not mistaken, nothing about what we should or should not do.

To establish that some proposed explicit definition of *right* corresponds to what is not mistaken to do, we must—somehow—show that that correspondence exists. Merely asserting the correspondence (by the inadvertently smuggled-in implicit definition of *right*, or by tricks with logical connectors) cannot suffice. We need an *argument* regarding what makes

6. As a practical matter, though, some terms' conventional, implicit connotations may be so strong that it would take an inordinate effort to avoid smuggling them in. In such a case, the least confusing approach is just to make the implicit definition the explicit one too.

something right (i.e., what makes something not-mistaken-to-do), not a mere stipulation by definition. (Chap. 7 below tries to outline such an argument.)

The same confusions arise with more-specific ethically laden terms, such as *ownership* (with regard to property rights) or *human life* (with regard to abortion). A supporter of libertarian capitalism may argue that you are morally entitled to use your own property for your exclusive benefit, because such entitlement is the very definition of the word *own*.

But by that definition, you have not established that anything is your own until you have (somehow) established that you are morally entitled to use it for your exclusive benefit. However, there is another definition of *own* that is often implicitly smuggled in—roughly, that if you have obtained an item by purchasing it, inheriting it, building it, and so forth, then you own it. Sleight-of-hand alternation between the explicit and implicit definition creates the illusion of having established that whatever you build, purchase, inherit, and so forth, you are necessarily entitled to use for your exclusive benefit. In abortion debates, to take another example, similar conflation occurs between *human life* in the sense of being an organism with a full DNA specification characteristic of our species, and *human life* in the sense of being morally entitled to protection from being destroyed.

Calling attention to the sleight of hand does not resolve the substantive question of whether or not each of the two conflated meanings does in fact imply the other. Rather, I am just reviewing some general ground rules for thinking about such questions, and for caution about how we express thoughts in words. The point so far is merely that the question of the correspondence between the two definitions does need to be addressed. The sleight-of-hand substitution of a smuggled-in definition for an explicitly stated definition gives the illusion of having established a correspondence that—whether it turns out to exist or not—has at any rate not been legitimately established by the sleight of hand.

Being on the lookout for smuggled-in definitions—in part by the diagnostic technique of asking whether it would seem to matter to use a different, made-up word with the same explicit definition—is an important aspect of any sensible, substantive investigation. (Similarly, if we decide for some reason to *change* the explicit definition of some word, we must take care to note that any statements previously established using the old definition now correspond to different propositions whose truth has not

necessarily been established.) We need to identify and eliminate semantic sleight-of-hand confusion before the serious discussion can even proceed.

Before the serious discussion does begin, there is one more procedural matter to address—a matter concerning the nature of paradoxes and their resolutions.

1.2.3 Paradoxes: When Arguments Collide

A paradox arises when two seemingly airtight arguments lead to contradictory conclusions—conclusions that cannot possibly both be true. It's similar to adding a set of numbers in a two-dimensional array and getting different answers depending on whether you sum up the rows first or the columns. Since the correct total must be the same either way, the difference shows that an error must have been made in at least one of the two sets of calculations. But it remains to discover at which step (or steps) an erroneous calculation occurred in either or both of the running sums.

There are two ways to rebut an argument. We might call them *countering* and *invalidating*.

- To counter an argument is to provide another argument that establishes the opposite conclusion.
- To invalidate an argument, we show that there is some step in that argument that simply does not follow from what precedes it (or we show that the argument's premises—the initial steps—are themselves false).

If an argument starts with true premises, and if every step in the argument does follow, then the argument's conclusion must be true. However, invalidating an argument—identifying an incorrect step somewhere—does not show that the argument's conclusion must be false. Rather, the invalidation merely removes that argument itself as a reason to think the conclusion true; the conclusion might still be true for other reasons. Therefore, to firmly rebut an argument whose conclusion is false, we must both invalidate the argument and also present a counterargument for the opposite conclusion.

In the case of a paradox, invalidating is especially important. Whichever of the contradictory conclusions is incorrect, we've already got an argument to counter it—that's what makes the matter a paradox in the first place! Piling on additional counterarguments may (or may not) lead to helpful insights, but the counterarguments themselves cannot suffice to

resolve the paradox. What we must also do is invalidate the argument for the false conclusion—that is, we must show how that argument contains one or more steps that do not follow.

Failing to recognize the need for invalidation can lead to frustratingly circular exchanges between proponents of the conflicting positions. One side responds to the other's argument with a counterargument, thinking it a sufficient rebuttal. The other side responds with a counter-counterargument—perhaps even a repetition of the original argument—thinking it an adequate rebuttal of the rebuttal. This cycle may persist indefinitely. With due attention to the need to invalidate as well as counter, we can interrupt the cycle and achieve a more productive discussion.

By way of illustration, here is a simple paradox, an inconsequential brain-teaser, that can serve as a warm-up exercise for trying to solve the other paradoxes discussed in this book. Also, despite its simplicity, this toy paradox has interesting structural parallels to some of the serious philosophical and scientific questions discussed in later chapters.

The paradox involves a curious property of ordinary mirrors: they swap an image's left and right, but not its top and bottom. For example, if you're wearing a wristwatch on your left hand, you see in the mirror a person who (construed as an actual person standing on the other side of a window) is wearing a wristwatch on the right hand (fig. 1.1). However, the person in the mirror does not appear to be standing upside down. Although the wristwatch is on the right hand instead of the left, the shoes are not on the head instead of the feet!

Thus, there is an apparent asymmetry between the mirror's treatment of the vertical dimension and its treatment of the horizontal dimension. And therein lies the seeming paradox, for we know that a mirror's reflection of a light beam is insensitive to which way is up. That is, the light's trajectory does not distinguish the horizontal from the vertical. How, then, can the overall result treat the horizontal differently from the vertical, if the result is made by components (the reflecting light beams) that have no such asymmetry?

We face here a paradoxical pair of arguments in support of opposite conclusions. One argument says that, given a set of events (the reflection of light beams) each of which treats the horizontal the same as the vertical,

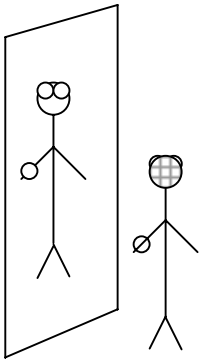


Figure 1.1
Your wristwatch is on your left hand, but your reflection's wristwatch is on its right hand.

the whole result (the appearance of the reflected image) must also treat horizontal and vertical the same. The other argument simply notes that in fact, the mirror does not treat the left–right axis the same as the up–down axis, since the reflected image's left and right are swapped, but not its top and bottom.

The problem is to resolve the contradiction between the principle that the mirror reflects light beams regardless of which way is up, and the observed fact that the resulting image shows left–right but not top–bottom swapping. In my anecdotal experience, it can take hours for smart people to correctly resolve this paradox. (If the problem is new to you, you may wish to try to solve it before reading on.) A number of tempting but incorrect resolutions come readily to mind, including the following three attempts to defuse the conflict between symmetric physics and the asymmetric perception of the reflection.

- *Rejecting physical symmetry* Maybe gravity really does somehow affect a photon's path. Doesn't general relativity say something about that? Perhaps the mirror somehow harnesses this effect to treat the image's vertical orientation differently from the horizontal.

- *Rejecting subjective appearances* Maybe the image's left and right aren't really swapped at all. Perhaps in reality, the reflection does match an unreflected person with a wristwatch still on the left hand. It may be just an illusion that it seems otherwise to us.

▪ *Rejecting reductionism* Perhaps there is a genuine (not just apparent) contradiction between the conclusion drawn from what the light beams do, and the observation of the actual image. That contradiction serves as a reductio ad absurdum of reductionism: the properties of the components (the bouncing light beams) simply cannot determine the orientation of the image made up of those beams; the image is irreducible to its components. The light beams and the image orientation must be understood on separate, incommensurable levels.

Before proceeding to the real answer, I would like to pause here for some observations that can be made even before that answer is at hand. I do so because in the case of several deep questions that this toy problem parallels, we have in fact been stuck for quite a while without a complete and convincing answer, so it helps to see how much can be done even without one.

Each of the first two attempted resolutions allies itself with one branch or other of the two conflicting arguments.

▪ Rejecting symmetric physics, we would postulate that the reflection's apparent asymmetry between horizontal and vertical in fact results from a corresponding underlying discrimination between horizontal and vertical in the very mechanism that produces the reflection.

▪ Rejecting subjective appearances, on the other hand, takes the opposite side: it stands by the horizontal–vertical symmetry at the lower level, and—to be compatible with that symmetry—denies that the emergent image swaps left and right at all, appearances to the contrary notwithstanding.

▪ Finally, the rejection of reductionism tries to uphold both branches of the contradiction by denying, in effect, that they can or must be reconciled: there simply is no fact of the matter as to which is the right one.

In fact, we have at hand the tools we need in order to be reasonably confident that these proposals are wrong, even before we know the right explanation.

Consider first the rejection of the underlying physical symmetry between horizontal and vertical. It is easy to see that even if the universe does provide for the gravitational diversion of photons, it might as well not for the purposes of this problem. To see why, you needn't even know that, as an empirical quantitative matter of general relativity, the gravitational influence on the paths of the reflecting photons is much too small to notice

here. Nor need you observe—though this is a good refutation in itself—that if you lie on your left side while looking in the mirror, you still see an image wearing its wristwatch on the other hand, but not wearing its shoes on its head, even though gravity is now pointing from your right to your left, not from your head to your feet. Hence, gravity doesn't explain the asymmetry. Even without these reasonable objections, there is a way to see that neither gravity, nor any other source of underlying physical asymmetry between the horizontal and vertical, could explain the paradox even if such underlying asymmetry existed.

To see this point, we need only observe that every aspect of the image's appearance—including its left–right swapping—is derivable from the standard optical model in which each light ray bounces from a mirror at the complement of the angle at which it strikes. From that model, you can deduce that the pattern of light beams reaching your eyes is the same as would be produced by a real person standing before you who looked just like you except for swapping left and right (fig. 1.2). Or—even without knowing the elementary optical principle describing the reflection of photons—you know that if any object is placed against a mirror, each part of the object that touches the mirror appears to touch the reflection of that very part. Thus, if you stand up against a mirror, with your nose, hands, and feet pressed against the glass, you see each of those body parts

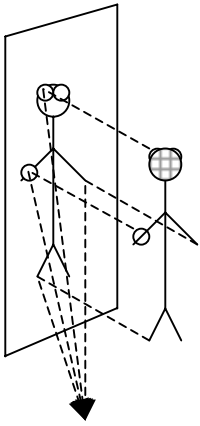


Figure 1.2
Light rays bouncing from your body reflect from the mirror, constructing a left–right swapped image.

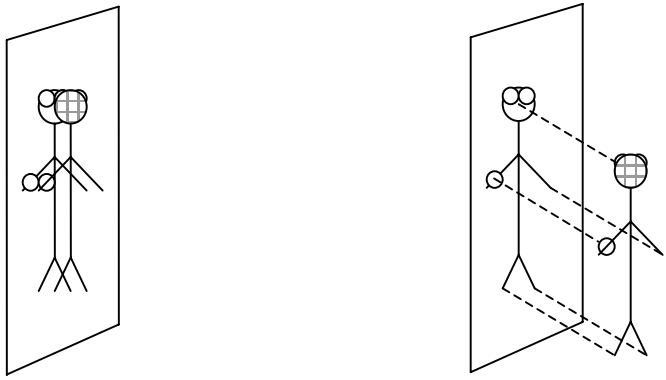


Figure 1.3
Each part that touches the mirror touches its own reflection.

touching its own reflection. And the resulting orientation of the reflection is maintained as you step back from the mirror (fig. 1.3).

Either by formal optical principles or by the touch-then-step-back method, you can *deduce* that the image has its watch on the other hand, but not its shoes on its head. And you know that the underlying model—whether the optical model or the touch-then-step-back model—is in fact symmetric with respect to its treatment of the horizontal and vertical directions. That is, *up* and *down* do not figure into the steps by which those models deduce what the image looks like.

Thus, you now know that any underlying physical deviation from horizontal–vertical symmetry would (even if it existed) be entirely beside the point, because you now know that even a symmetric underlying mechanism somehow automatically gives rise to the apparently asymmetrically swapped image! Even if the real world had some underlying horizontal–vertical physical asymmetry in the physics of photons, that asymmetry could not address the paradox of how a fully symmetric model could still produce the same mysterious higher-level asymmetry.

And conversely, if we resolved the paradox and understood how that high-level asymmetry could be produced from symmetric underpinnings, we would no longer need to appeal to any underlying physical asymmetry to explain the reflection's orientation. This consideration does not, in itself, disprove the physical-asymmetry hypothesis. It does, however, render the

hypothesis entirely *superfluous* to the present paradox. So at the very least, the reasons the paradox might motivate us to entertain the physical-asymmetry hypothesis are shown by the above analysis to be misguided, because a version of the paradox remains in full force even if we grant the phenomena postulated in the hypothesized explanation.

Thus, despite its superficial appeal, the rejection of underlying physical symmetry is fundamentally misguided as an attempt to resolve the paradox; it is a superfluous hypothesis. On the other hand, the rejection of subjective appearances, or the rejection of reductionism, are not so much explanations as acts of desperation. Rejecting appearances blatantly denies a plainly demonstrable fact, and in any case is felled by the argument just given: the asymmetrically altered orientation of the image's appearance is readily deducible from the (symmetric) underlying principles, and so is not just illusory (unless our deductive apparatus, as well as our perceptual apparatus, falls prey to the illusion). And the antireductionist dodge, far from answering the question, insists instead that it is unanswerable just because the answer is not yet available.

But rejecting the above proposed resolutions of the paradox still leaves the paradox unresolved. We know there is an underlying mechanism that treats up–down and left–right symmetrically. Therefore, it seems that the image can swap the original's left and right if and only if the image also swaps the original's top and bottom. But this argument must be mistaken somehow, because the image—unlike the original—wears its wristwatch on its right hand, swapping left and right. But—like the original—it wears its shoes on its feet, not its head, preserving up and down.

Resolving a paradox requires not just convincing ourselves that one of the conflicting arguments must be mistaken (in this case, the symmetry argument that says either both, or neither, of the two dimensions must end up swapped), but also requires showing just what is wrong with that argument—if it gets to the wrong conclusion, at what point does the argument go awry? In addition to countering the bogus argument (by arguing for the opposite conclusion), we need to invalidate the argument (by showing which step does not properly follow).

At this point, the facts of the underlying model or mechanism are no longer in question (or at least no longer relevantly in question—whatever minute

effect gravity may have on photons, we know it is beside the point here). We know, even before solving the problem, that we have at our disposal the means to derive all the relevant details of the problem, both at the underlying level (using optical principles or just the touch-and-step-back method) and at the emergent level (the reflection's apparent orientation).

Thus, all the pieces we need are in our hands. What remains is to interpret them in such a way as to fit them together. Noticing that an asymmetric swapping is derivable from a symmetric underlying model tells us not only where not to look for the answer (i.e., supposedly asymmetric physics), but also hints strongly at were we *should* look. In particular, let us look at the details of that derivation and see at what point the asymmetry creeps in.

Imagine a rectangular box, each of its six sides with a distinct appearance, held before a vertical mirror. There are no words or pictures on the box, so we might construe it to be in any of several orientations: right-side up, upside down, on its side or back, and so on. Let us designate its ''front'' to be the surface facing the mirror—the gray surface in figure 1.4. And let us call the gray surface of the reflected image *its* front. If the front of the real box faces north, the front of the reflected image faces south. But the image's four sides perpendicular to the front each face the same way as the like-colored side of the real box (up, down, east, or west). Unsurpris-

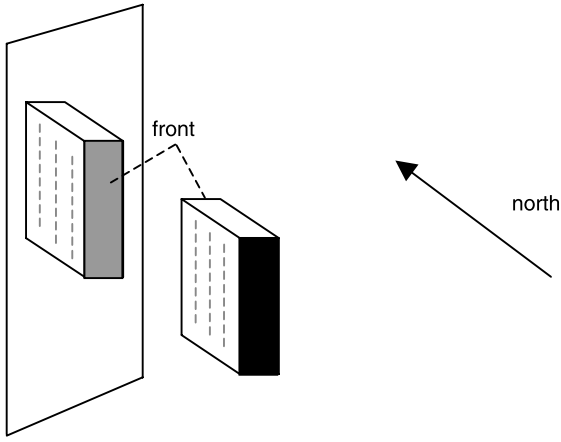


Figure 1.4
A mirror reverses the direction that the front side faces, but not the other directions.

ingly, the mirror reverses the directions of front and back, but not the other four directions.

Suppose next we choose the upward-pointing side—the white surface—of the real box and call it the ''top.'' And suppose we call the white surface of the reflected box the top, too. Specifying both the front side and the top side then tells us which the right and left sides must be. But then the furrowed side is the left side of the real box, but is the right side of the reflected box. Here, the apparent asymmetry between horizontal and vertical reasserts itself: left and right are swapped (fig. 1.5).

But suppose that instead of choosing a side to call the ''top,'' we had chosen a side of the real box to call ''left''—in particular, the furrowed side—and then stuck to that designation for the furrowed side of the reflected box as well, as in figure 1.6. Specifying the front side and the left side tells us which must be the top side and the bottom side—but then the color that is the ''top'' of the real box is at the ''bottom'' of the reflected image, and vice versa! Thus, if we choose to designate the furrowed side ''left,'' and derive top–bottom from that, rather than designating the white side ''top'' and deriving left–right, then the mirror is performing top–bottom swapping instead of left–right swapping. Accordingly, the reflected image is upside-down.

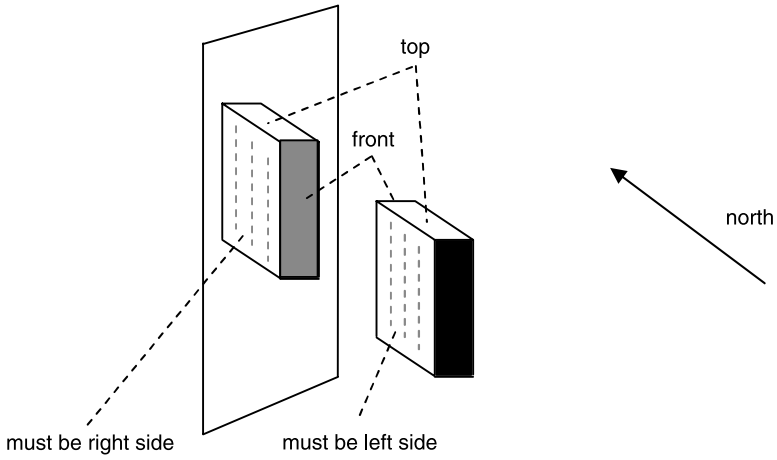


Figure 1.5
Specifying the front and top sides tells us which the right and left must be.

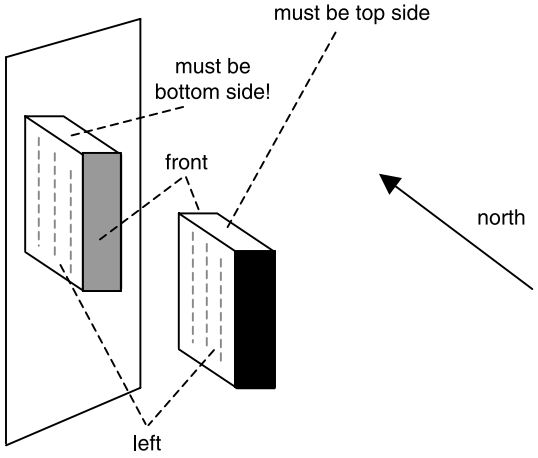


Figure 1.6
Specifying the front and left sides tells us which the top and bottom must be.

At last the resolution of the paradox becomes comprehensible:

- The mirror treats only the perpendicular dimension differently from the others—the surface facing the mirror faces north, but its reflection faces south.
- If we insist on construing the reflection of the front as still the front, and the reflection of the top as still the top, then it follows that the reflection's right and left are swapped.
- But if instead we identify front with front and left with left, then it follows that the reflection's top and bottom are swapped instead.

Thus, it is not the mirror itself, but rather our choice of which of those identifications to emphasize, that determines whether there seems to be horizontal or vertical swapping.

Presumably, people's approximate left–right symmetry—in contrast with our conspicuous top–bottom and front–back asymmetry—induces a preference for preserving the identity of the top and front rather than the identity of the left or right. After all, it is much easier to construe the reflection of a right hand as a transformed left hand than to construe the reflection of a head as a transformed pair of feet. A blank box with differently colored sides does not induce as strong a preference, and thus makes it apparent that the opposite interpretation can be made instead.

Accordingly, the attempted resolution by rejecting subjective appearances was *almost* correct: there is an illusion of sorts about the image's swapping of sides. But the illusion is not that the image seems to have any swap at all. In fact, the swapping is quite real—there is no way to reorient an actual object to make it look just like the object depicted by its reflection. Rather, the object would have to be taken apart and reassembled, swapping one surface for another. The illusion is just that the choice of there being a left–right swap, rather than a top–down swap, *seems* to be inherent in the image. But in fact, the image can be described either way, and it is a psychological preference unrelated to the optics of mirrors that chooses one interpretation over the other.

The swapping per se is therefore not illusory—it was merely misinterpreted. In retrospect, we should have known that the mirror's horizontal–vertical reversal asymmetry must be induced somehow by the psychology of the interpretation of the image. Given the deducibility of the image from the symmetric underlying model, there's simply nowhere else for the asymmetry to have come from.

Thus, of the two conflicting arguments at the outset, the correct one observes that the mirror must treat the left–right axis the same as the top–bottom axis. The subtly flawed argument says that the mirror does not in fact treat those axes alike, since right and left swap, but not top and bottom. But that conclusion does not follow. For although the reflection is genuinely transformed, its transformation can correctly be described as either a left–right swap *or* a top–bottom swap—just not both at once. That ambiguity is compatible with the underlying mechanism's symmetry (since claiming either swap turns out to be just as valid as claiming the other), but still allows for exactly one of the two dimensions to be swapped.

Of course, the mirror paradox is too trivial for any of the spurious resolutions above to be taken seriously for long. But if the paradox were much harder and more important to us, then dedicated schools of thought might form behind each of the incorrect resolutions: one school revising its view of the physics of the universe to accommodate the perceived asymmetry, another just denying or ignoring the inexplicable perception, and still another accepting the paradox as an eternal mystery, abandoning the possibility of a coherent factual resolution. After all, what else can we do in the

face of a seemingly intractable conflict between subjective appearances and known underlying processes?

That, I maintain, is the situation today with many serious, nontrivial problems (adjudicating between choice and determinism, between quantum mechanics and observer-independent objectivity, between the apparent forward flow of time and the time symmetry of physical laws, between a moral and a mechanical view of the universe . . .). And the question about what to do in the interim, pending a convincing resolution, is not merely rhetorical. We can, alternatively, expect that there must turn out to be a reasonable answer—one that is consistent both with subjective appearances, and with seemingly incompatible underlying processes—even if the details of the resolution have yet to be worked out.

1.3 Right Side Up (How to Read This Book)

To a casual observer, the earth seems flat. I remember wondering, as a small child, what the edge of the earth looked like (I tentatively envisioned a long stone wall beside a parking area and gift shop). Of course I'd been told the earth is round, but I took that to mean circular, not spherical.

Europeans contemplating the roundness of the earth are said to have remarked that people in China must then stand on their heads. Such confusion is understandable. Abandoning the idea of a flat earth is not just a matter of geometry; it also involves revising our notions of physics. With a seemingly flat earth, the directions *up* and *down*—the directions from which and to which things fall—appear to be constant no matter where you are. But if the earth is round, and if people on the other side walk on the ground just as we do, then *up* and *down* must work differently than a flat-earther would think—they must point from and toward the center of the earth (respectively), so that *down* points in opposite directions from opposite sides of the earth. Here, unlike with the up–down considerations in the mirror-asymmetry paradox, gravity does form a critical part of the explanation.

The ramifications of a round earth illustrate a basic principle of rational inquiry: ideas intertwine, and their entanglement can impede the revision of one false idea unless and until we are prepared to entertain the possibility of having to abandon many other beliefs as well. Given a network of mutually supporting fallacies, any narrow effort at reform is doomed, be-

cause correcting any one fallacy in isolation leads to apparently absurd contradictions with the other false notions—often ones that bear on our direct personal experiences, as with people seemingly having to stand on their heads. We cannot escape the apparent absurdities until we correct the other fallacies as well.

Yet we cannot overcome all such intertwined errors simultaneously, of course. What is needed instead is a willingness to tentatively tolerate the apparently absurd consequences of a new idea until we have had a chance to investigate those consequences more carefully. Eventually, after each such consequence has been explored in turn, we can look back and ask whether any of the mutually supporting old notions are still standing.

So it is with the ideas in this book. The concept of a mechanical, deterministic universe seemingly clashes with important intuitions about consciousness, choice, ethical responsibility, and other matters. And nowhere is the entangled confusion greater than in the realm of quantum mechanics, where physical evidence itself seems to argue for an observer-dependent, hence nonobjective world (although, as I argue in chapter 4, that supposed physical evidence is being misinterpreted because of an intertwined misconception concerning conscious observation). To show that a completely mechanical universe does not, so to speak, have us standing on our heads—does not have us devoid of consciousness, choice making, a foundation for ethics, or an explanation for the results of the quantum EPR experiment—we must carefully explore how those deeply important phenomena relate to a mechanical reality.

Although the brunt of this book is to argue for the mechanical view as a coherent whole, what follows can also be taken as a series of vignettes of varying length and depth, among which one may pick and choose those of interest. The only caveat is that, as just noted, a piecemeal selection will sever some important connections and leave them dangling.

Some parts of this book are easier to read than others. The book requires no more than a high-school math and science background, but some chapters (especially 4, 5, and 6) plunge into technical details. Many readers will prefer to skip the more detailed sections, and I've tried to write the book so that the rest of it still makes some sense.

But I'm convinced that the detailed parts may be useful even to readers who mostly page past them, because the pages at least convey a sense of how much detail does underly the less technical face of the presentation.

Many works of popularized science, I fear, provide oversimplified glosses of technical, mathematical principles, leaving readers with a false impression of having grasped the idea. Not knowing that the real content was omitted, they may end up further from the truth than before they started reading. Whatever other mistakes I may make here, I have tried to avoid that one—instead, I try to present both a summary view and a detailed view, inviting readers to sample from the latter at their leisure.

The book proceeds as follows:

- Chapter 2 explores how inanimate, mechanical matter could be conscious, just by virtue of being organized to perform the right kind of computation.
- Chapter 3 explains why conscious beings would experience an apparent inexorable forward flow of time, even in a universe whose physical principles are time-symmetric and have no such flow, with everything sitting statically in spacetime.
- Chapter 4, following Everett, looks closely at the paradoxes of quantum mechanics, showing how some theorists came to conclude—mistakenly, I argue—that consciousness is part of the story of quantum phenomena, or vice versa. Chapter 4 also shows how quantum phenomena are consistent with determinism (even though so-called hidden-variable theories of quantum determinism are provably wrong).
- Chapter 5 examines in detail how it can be that we make genuine choices even in a mechanical, deterministic universe.
- Chapter 6 analyzes Newcomb's Problem, a startling paradox that elicits some counterintuitive conclusions about choice and causality.
- Chapter 7 considers how our choices can have a moral component—that is, how even a mechanical, deterministic universe can provide a basis for distinguishing right from wrong.
- Chapter 8 wraps up the presentation and touches briefly on some concluding metaphysical questions.

Thus, the three main parts consist of chapter 2 (consciousness), chapters 3 and 4 (physics: time and quantum), and chapters 5 through 7 (choice and ethics). Chapters 5 through 7 present the book's most original technical content; the other chapters largely review familiar ground, albeit with some expository innovations.

2 Dust to Lust: How Groups of Atoms Can Think and Feel

We see and hear and touch the world around us—but so do cameras, microphones, and probes. The difference, though, is that we can *know* we're doing those things. We're aware of our perceptions; our sights and sounds are *conscious*.

We move our arms and legs and speak words. But so do robots or tape recorders. The difference is that we are free to choose what to do. Machines just display rigid, preconstrained behavior.

And unlike any inanimate object, we have thoughts and feelings flowing through our minds. We can recall events from this morning, or years ago. We can speak silently to ourselves, conjure up mental images of things we're not actually looking at, or be aware of things we believe to be true, or of things we want to be true, or of things we plan to do. We feel love, anger, joy, sorrow, excitement, boredom. All this happens (primarily) inside our heads. We know of its occurrence, but not the way we know of physical objects or events—we do not see, hear, touch, taste, or smell our thoughts and our emotions. Compared to hard, tangible, physical objects—or even compared to the invisible air, which at least we feel when it moves—the events in our minds seem to have a separate, ghostlike kind of existence, distinct from the physical realm.

Most importantly, these nonphysical-seeming thoughts and feelings seem to matter intrinsically, in ways that inanimate, insensate objects and events do not. If a machine breaks or ceases to exist, it makes no difference to the machine, because the machine neither knew nor cared that it existed in the first place. Your consciousness gives you value—to yourself and also to others, who have certain moral obligations toward you, as you have toward them. No one has an obligation toward a machine, nor can a machine itself have moral obligations. Or so it appears.

Yet the more we learn about the world *outside* of our minds, the more mechanical it looks. Even thousands of years ago, it was obvious that the sun, moon, and stars follow rigid, predictable patterns. The sun rises and sets repeatedly, and its path across the sky varies cyclically, in tune with the seasons. The stars rise and set almost in tune with the sun. The moon's journey follows a separate schedule, but the shape of its crescent is tied in part to the position of the sun. The paths of the planets are more complex, but still regular enough that we can predict, many years in advance, where they will be.

Just several hundred years ago, humanity discovered that, contrary to appearances, the sun, stars, and planets do not orbit the earth—it just looks that way from here. More remarkably, it turned out that there are precise principles that describe the motion of those bodies, and they are the same principles that describe the motion of the things around us: rocks, water, air, and all the rest. As science flourished, we found that all matter and energy is composed of elementary particles. First we discovered atoms and molecules. Then, inside those, we found protons, electrons, and neutrons. Then, inside *those* were quarks and leptons (and even further inside, we suspect, are many-dimensional strings)—entities that combine to form all that exists in the world.

Most remarkably of all, each of the ever-smaller building blocks we find also behaves in a way that can be described by rigid, mechanical, mathematically expressible principles. There are just a few kinds of building blocks, and each kind always responds the same way to the things around it. We haven't yet pinned down the smallest building blocks and the corresponding set of rules, though many physicists suspect we are getting close.

Recall the mention in section 1.1 above of Laplace's clockwork paradigm. According to that paradigm in its original form, objects in the universe have smooth, continuous motion that conforms to mechanical, mathematical rules. In Fredkin's or Wolfram's variants, the universe is instead a machine known as a cellular automaton, with an enormous number of tiny, discrete cells—like the squares of a chessboard—each of which is in one of several discrete states at any moment—analogous to having one of several kinds of chess pieces (or none at all) on it. Cellular automaton rules specify what new state a cell will acquire, depending on its current state and its neighbors' current states.

Given either sort of mechanical paradigm, continuous or discrete, the paradoxes blossom. Our own bodies and brains turn out to comprise the same kind of elementary particles as the rest of the world. If everything in the world—including you and me—is just an assembly of purely mechanical building blocks, then all such things (including you and me) are themselves just machines. But then what about our ghostlike thoughts and feelings, our consciousness?

2.1 The Case against Ghosts

One prominent notion is that we have both a ghostlike component (our consciousness or soul) and a mechanical component (everything else, including our body). The mechanical component is governed by the usual physical laws. The ghostlike component, unconstrained by those laws, can be said to be *extraphysical*. That is, the ghostlike component is something in addition to the kinds of things that exist in the physical realm, something ontologically extra.[1] This so-called *dualist* view was advanced by Descartes in the 1600s.

Dualism is a tempting compromise, but an awkward one, for reasons that are well known. The problem is that the mechanical principles that govern each particle of our bodies (and of the things around us) already specify how each of those particles behaves, which in turn specifies how each of us behaves as a whole. But in that case, there is no room for the ghostlike component to have any influence—if it did so, it would have to make some of the particles sometimes violate the principles that all particles are always observed to obey whenever we check carefully. (Descartes was admirably precise about the locus of this supposed intervention—he proposed that the interface between the ghostlike component and the physical world occurs within the brain in the pineal gland.)[2] Thus, we have the *mind–body* problem: how can we reconcile the nature of the mind with the mechanical nature of the body?

Some see quantum-mechanical uncertainty as the wiggle room that could let a ghostlike consciousness nudge some of the particles in our body

1. *Ontological* considerations are concerned with what kinds of things exist.

2. There are many excellent discussions of Descartes's dualism from a modern point of view, e.g., in Dennett 1991.

without violating the rules of physics. But in fact—even apart from the newer, deterministic interpretation of quantum mechanics discussed in chapter 4—any such nudging would at least constitute a change in the probability distribution for some of the particles in our body, and even that would break the (probabilistic) rules that particles always seem to obey.

Granted, it *could* be the case that particles somewhere in our brains behave differently than particles ever do when we watch them carefully, violating otherwise exceptionless rules (be they deterministic or probabilistic rules). But since the rules *are* otherwise exceptionless (as far as we can tell), there should be a strong presumption that there's no exception in our brains either—especially in view of the longstanding retreat of other beliefs about the alleged physically exceptional behavior of conscious or living organisms. The doctrine of *vitalism*, for instance, supposed that there is some distinctive ''life force'' that animates living things, enabling them to grow and move. But the more we learned of biochemistry—DNA and RNA, ATP energy cycles, neurotransmitters, and the like—the more we understood that the growth and movement of living things is explicable in terms of the same molecular building blocks, following the same exceptionless rules, as when those building blocks exist outside of animate objects.

And the more we learn about computation and neuroscience, the more we discover how cognitive processes that were once supposed to require an ethereal spirit—perception, motor control, memory, spatial reasoning, even key aspects of more general reasoning (e.g., deduction, induction, planning)—can be implemented by basic switching elements (e.g., neurons or transistors) that need not themselves be conscious, or even animate. By monitoring brain activity, we can see different regions of the brain performing computations when different sorts of cognitive functions are performed (language, singing, spatial imaging, etc.). And when certain brain regions are damaged by injury or illness, the corresponding cognitive abilities degrade or vanish. To be sure, we are still far from understanding human cognition as a whole. But the trend in our knowledge does not lend comfort to the expectation that any particles in our brain will, at long last, ever be found to deviate sometimes from the same rules that such particles otherwise always obey.

I do not attempt here to review the wealth of evidence that particles behave the same (mechanical) way within our bodies as they do elsewhere,

and that mechanical building blocks can be assembled in such a way that the assembly displays the sort of behavior we see in living beings, or even in intelligent living beings. Instead, I merely want to argue here for the *coherence* of that view by identifying and resolving some key apparent contradictions that the view leads to. If these seeming contradictions are cleared up, we will have no more reason to suspect that our minds include a nonmechanical, extraphysical component than we have to suspect that rocks or toasters have such a component.

Although our thoughts and feelings don't *seem* like mechanical processes, we know there is much about our minds and brains that is not as it seems. It doesn't seem like the visual images we receive are upside-down on our retinas; yet we know they are. It doesn't seem like moving or feeling our limbs is accomplished indirectly by sending electrical signals through our spinal cords; yet we know what happens if that pathway is severed. When mechanical intermediary states are simply hidden from scrutiny, it seems like they're not there. If we could watch our neurons in detail while we're thinking, and see the detailed correspondences, then our thoughts might indeed seem just like neural events (though seeming that way would not itself prove that our thoughts *are* neural events—just as seeming the other way does not prove that our thoughts are *not* just neural events).

As just argued, any ghostlike component does not plausibly override the mechanical rules according to which our bodies' and brains' particles behave. Hence it does not influence those particles. But *unless* a ghostlike component influences some particles, its decisions (if it is indeed the part of you that makes decisions) have no effect on our bodies' actions.

Leibniz, in his seventeenth-century theory of *monads*, proposed that mind and the nonmental world, though lacking any influence on one another, remain synchronized by a "preestablished harmony." Each is independently rigged so that its entire future unfolds in a way that will mesh with the other, just as two carefully contrived voice recordings—each made separately from the other—can be played back simultaneously to give the illusion of a conversation taking place between the two recorded parties, though in fact there is no interaction.

But if supposedly nonmechanical extraphysical consciousness perfectly mirrors what mechanical atoms do, then the behavior of consciousness

itself is just as mechanically constrained—it too then turns out to be just a machine. But if extraphysical consciousness is merely redundant, mechanically duplicating part of what happens in the ordinary physical world, then there is no reason even to suspect that this extraphysical component is present.

A similar objection applies to attempts to turn the tables by suggesting that although the ghostlike component does not influence the physical brain and body, there is influence that runs in the other direction. That is, our decisions—and our memories, and reasoning, and emotions, and other cognitive abilities and phenomena—are indeed implemented mechanically, computationally, by our neurons. Still, by this proposal, there is something else: an extraphysical consciousness that is somehow *generated* by the physical, computational events in your brain. Without that consciousness, your body would still behave as it always does—the difference would be externally imperceptible. But internally, there would be no conscious *you*—no sensation of experiencing anything, no awareness of thoughts, feelings, perceptions, and so on. It would be like being a rock—or so the story goes.

With this variant of dualism, the supposed generation of consciousness from physical brain processes does account for the mind–body alignment without requiring either a preestablished synchronization, or an inexplicable massive coincidence. But there remains a question of how you could then have any awareness of your consciousness. If it is the mechanical component of your mind that implements the details of your thinking and knowing and perceiving and deciding and acting and remembering, and that generates your consciousness as a side effect, with no influence in the opposite direction, then in particular anything you think or say about your consciousness cannot be influenced by any aspect of your consciousness itself—it cannot even be influenced by your consciousness's very existence. In that case, when you say things like "But I can *feel* that I have consciousness. It does not feel like there's just electrochemical activity taking place in my brain," your saying or thinking that could not be in any way influenced or explained by your extraphysical consciousness, even if your extraphysical consciousness exists.

It's not just that you'd look the same to others—making that same remark, for example—even if you actually had no extraphysical consciousness. You would indeed look and act the same, of course. But more than

that, you'd be making that same remark as a result of the same causes that produce the remark if you *do* have an extraphysical consciousness—namely, the same physical, mechanical, computational events in your brain. Even if your extraphysical consciousness were real, the apparent perception of it by your physical brain would still be illusory (if the extraphysical stuff is generated by, but does not influence, the brain's configuration of physical particles). The apparent perception would have to be consistent with and explicable by the usual physical regularities, and thus not be due in any way to what is supposedly perceived. If you would perceive (and report and remember) the same thing with or without the existence of extraphysical consciousness, then that perception (or report or memory) constitutes no evidence for extraphysical consciousness.

Thus, if extraphysical consciousness is not something that physically affects any of the particles in our brains and bodies, we must choose either the view that consciousness is indeed extraphysical—in which case we cannot even perceive it—or the view that we can and do perceive and remember and report it, in which case it must not be extraphysical. That pair of alternatives—although not flatly disproving the existence of extraphysical consciousness—certainly vitiates any motivation, on the basis of the seeming perception of it, to think it does exist. The hypothesis of its existence becomes superfluous with respect to the seeming perception (just as, in sec. 1.2.3's mirror paradox, discovering that the optics or press-and-retreat model already predicts the seemingly asymmetric swap renders the underlying-physical-asymmetry hypothesis superfluous, unmotivated).

Of course, there is *something* we perceive when we think we perceive our extraphysical consciousness—that is, when we perceive conscious events such as our thoughts, feelings, perceptions, and so on. Such events affect our words and behavior (as well as affecting one another). These events are no mere illusions, just as the left–right swap in the mirror paradox could not have been simply an illusion—although a key aspect of its interpretation was indeed mistaken and illusory, as too with consciousness. If we define *consciousness* ostensively[3]—as whatever it is we are perceiving when we perceive those cognitive events—then we can ask whether consciousness turns out to be a ghostlike, extraphysical phenomenon, or whether on the

3. A term is defined *ostensively* when we cite exemplars of it and define the term as things that are *like those*. (The word is not to be confused with *ostensibly*.)

contrary it is just a particular physical, mechanical property or process of our brains. This is a substantive empirical question.

Alternatively, of course, we could simply *define* consciousness as something extraphysical. But doing so would not, of course, make our thoughts, feelings, and so forth turn out to be extraphysical. Rather, with this alternative definition, to establish that we perceive our own consciousness in the first place, we would have to establish not only that we perceive our thoughts, feelings, and so forth (which is obvious), but also that what we are perceiving when we perceive those things is indeed something extraphysical. Thus, as discussed in section 1.2.2, we cannot resolve a substantive question by trying to build a favored answer into a definition. The question of whether our thoughts, feelings, and so on, are extraphysical or not—whether they are ghostlike or mechanical—is an empirical question, to be established by evidence and arguments, not by issuing definitions. The most convenient definition to use for this inquiry is the ostensive one, which is neutral as to whether what we are referring to turns out to be extraphysical.

In sum, there is no evidence for a ghostlike, extraphysical consciousness that somehow alters the trajectory of some of the particles in our brains (or that is generated by or synchronized with those particles, without influencing them in turn). The entire history of scientific knowledge has shown a steady retreat from the superficially seductive notion that life, or human life, or consciousness, is physically special, powered by some vital force that is not subject to the principles that govern inanimate particles. Instead, it looks more and more as though life—and human life, and human thought—consist of special *arrangements* of ordinary particles following ordinary rules.

Just as a handful of different CPU instructions (or different logic gates) in a digital computer can be combined in exponentially many ways to produce an endless variety of computer software (word processors, Web browsers, email readers, DVD players, spreadsheets, flight simulators, fantasy games, viruses, firewalls), and just as the twenty-six letters of the English alphabet can be combined in exponentially many ways to produce an endless variety of written works, so too can a small number of kinds of particles be combined in many ways to implement every object that exists, animate or inanimate. It's not the building blocks that are special to each

kind of object. What's special instead is a particular *organization* of the same old building blocks.

But why, then, does our consciousness not feel to us like something that just consists of ordinary atoms, or (looking at it at a higher level) of neurons and electrochemical activity? Why, indeed, does it feel like *anything* to us, if it is just a special arrangement of inanimate particles? Why doesn't the brain just do what it does, computing and reacting and manipulating our bodies, without experiencing anything? If each particular conscious thought is just a physical or computational event, then why instead couldn't that same event (with the same physical properties) occur without its being a conscious event? To address these questions, we need at least an outline of how there could be a mechanism of consciousness.

2.2 Cartesian Camcorders, Big Red Rock-Eaters, and the Light in the Refrigerator

The human brain is astonishingly complex and multifaceted.

- It has distinct specialized systems for seeing, hearing, touching; for walking and for handling objects; for spatial reasoning, associating words with meanings, combining words to form sentences; for understanding aspects of causality; and perhaps for various specialized domains of knowledge: social, physical, musical, possibly even botanical (see Steven Pinker's *How the Mind Works* [1997], for example).
- The brain understands and predicts aspects of the world around us. It anticipates things that happen independently of its own choices, and it makes choices in expectation of—and for the sake of—their outcomes.
- The brain's choices are influenced in part by its emotions (affection or anger, for example), by its sensual preferences (pursuing the taste of food, comfortable temperatures, orgasms, and so on, while avoiding painful collisions and falls, contact with red-hot surfaces, and the like), and by its more abstract preferences (for companionship, for knowledge, for aesthetic stimulation, for acquiring skills).
- The brain remembers things—in the form of learned facts and acquired skills and experience-based anticipations, and in the form of *episodic* memories—videolike recollections of past events, thoughts, and experiences (although the brain's videolike record is neither exhaustive nor error free).

Especially in light of the brain's intricacy—and correspondingly, that of the mind—consciousness is impressive for its seemingly seamless integration of disparate mental phenomena. There appears to be a single, unified *you*, a something (or a someone) that is aware of, that has a view of, the memories, knowledge, skills, perceptions, thoughts, and feelings that arise in your mind. And it is seemingly by virtue of being so viewed that those phenomena are conscious, rather than inanimate events unfolding unnoticed. When you look at a book in your hands, you *know* you are doing so. Unlike a camcorder, you do not just mechanically record the event.

In *Consciousness Explained* (1991) and *Freedom Evolves* (2003), Daniel Dennett describes your apparent view of the panoply of your mind in terms of a *Cartesian Theater*. A conscious *you*, metaphorically perched in the audience, watches (and perhaps directs) the cognitive events that unfold on the stage. The name *Cartesian* refers to Descartes's dualism, although the concept applies even if the metaphorical audience member is a purely physical, computational entity.

Dennett marshals modern neurophysiological evidence to cast doubt upon the Cartesian Theater idea—even in its nondualist form. He argues that there is a limit to the precision with which you can meaningfully specify the time and place where, say, an image from your retina passes into your consciousness. You cannot pin down the millisecond when it occurs, or the cubic millimeter of brain matter where it occurs, because there is no such precise time and place. Rather, the physical processes that implement consciousness—even the specific consciousness of a specific event—are sprawled over many simultaneously operating modules of the brain, and their operation takes at least a substantial fraction of a second.

Traditions as far back as Buddhism have challenged the notion of a central conscious observer. The Buddhist doctrine of *anatta* speaks of a stream of thoughts or experiences which themselves constitute consciousness, but with no separate, nontransitory entity *having* or *viewing* these experiences.[4] Indeed, from an engineering standpoint, what need is there for a separate observer—why do the mental events not just unfold unobserved? Or if a

4. This doctrine's correspondence to aspects of a contemporary computational view of mind is among the topics explored in *Gentle Bridges* (Hayward and Varela 1992), which documents an exchange of ideas between the Dalai Lama Tenzin Gyatso and Western cognitive scientists at the 1987 Mind and Life Conference.

mental event's consciousness does consist of the event's being observed by an internal observer, then the observer must itself be observed, or we could not be conscious of our consciousness (or of the observer's decisions in its directorial capacity, if any). Does consciousness then require an infinite regress of observers observing other observers? That would be implausibly extravagant, especially if a finite physical brain implements consciousness. But the apparent alternative—that some events in the brain are somehow just intrinsically conscious, intrinsically self-aware, rather than requiring a separate observer—has difficulties as well. What could make something intrinsically self-aware?

Our conscious mental experiences do *feel* intrinsically self-aware. When, say, I am conscious of a flower I'm looking at, I cannot seem to dissociate my perception of the flower from my awareness that I perceive the flower—they feel like two aspects of the same conscious phenomenon. Being aware of the flower, but not of the very awareness, would seem more like the status of an unconscious computer program that, hooked up to a camera, analyzes the incoming image and classifies it as a flower, but doesn't know it does so, feeling no experience of doing so. Similarly for my consciousness of purely internal events, like silently recalling a tune or thinking what my next move might be in a chess game.

The very fact that (or even the mere logical possibility that) a computer program *can* recognize a flower without knowing or experiencing that it is doing so (indeed, without experiencing *anything*) illustrates that the recognition per se is not intrinsically self-aware. Rather, we can coherently conceive of seeing and recognizing a flower unconsciously. Indeed, each of us sometimes has that kind of unconscious awareness—for instance, if I am walking across a lawn engrossed in a conversation, I might change my trajectory to avoid stepping on a flower in my path (or at least, I might reasonably surmise that that was why I stepped around the flower, if I were later to watch a video of my steps) even though at the time, my attention was directed elsewhere and I was not aware of noticing the flower's presence. In that sense, I was not *conscious* of the flower's presence, even if the video suggests that part of my brain did act from a recognition that the flower was there.

Still, even though recognition of a flower can be unconscious, at other times when it *is* conscious, its consciousness does not feel like a dissociable aspect of the recognition, like something that could be taken away to leave

only the unconscious recognition. By analogy, if there is a rock of some particular shape, the shape is not something you can subtract from the rock, leaving just a shapeless rock. In that sense, the particular shape is intrinsic to the particular rock—even though another rock can be just as much a rock but not have that shape. Likewise with the consciousness status of a particular perception or thought—or that, at any rate, is how it feels.

I argue, though, that the conscious experience that (usually) accompanies, say, my brain's recognition of a flower is indeed a phenomenon *in addition* to my brain's recognition of the flower—the consciousness is not intrinsic. The consciousness is a phenomenon that could indeed be subtracted away, leaving just the (unconscious) recognition—despite my impression that there is a single, irreducible event that somehow has self-awareness—conscious experience—as an intrinsic property. What, then, could account for both this impression of irreducibility and its falsehood? And what could provide for the self-awareness (and awareness of that self-awareness, and so on) without requiring an infinite series of observers, observers of observers, and so on?

Here is a proposal. Each cognitive event—such as recognizing a flower—is itself devoid of consciousness. But there is an internal memory system—what we might call a *Cartesian Camcorder*, in tribute to Dennett's Cartesian Theater metaphor—that records a stream of selected mental events—perceptions, thoughts, and so on—that the machinery deems salient. A recorded event can be played back and ''watched'' by other parts of the cognitive system, either immediately or much later; this process turns out to be what consciousness consists of. The event of ''watching'' the playback is itself a cognitive event that can be recorded and played back (just as, with a literal camcorder, you can shoot a video of yourself watching a previous video).

What I call the Cartesian Camcorder view has a different emphasis than, but is consistent with, Dennett's *multiple-drafts* view of consciousness. (The camcorder metaphor, as elaborated below, also distills aspects of, e.g., the *self-model* theory of Metzinger [2003], and the *global-workspace* model of Baars [1988].) Drawing from a number of technical speculations about cognitive architecture (prominently including Marvin Minsky's *The Society of Mind* [1986], and Baars), Dennett offers the multiple-drafts view as an al-

ternative to the Cartesian Theater. According to the multiple-drafts view, various loosely coordinated versions (or drafts) of the recording are edited by separate, interacting processes, but there is no central, definitive version.

That's okay—the Cartesian Camcorder recording I speak of may well be implemented, if you look inside the machinery, by a collection of somewhat separately maintained versions. Whereas Dennett focuses (quite reasonably) on ways in which consciousness is *not* like the Cartesian Theater model (if you look at consciousness microscopically enough), I am more concerned here to stress ways in which consciousness *is* like the Cartesian Theater model, if you *don't* look at it too microscopically—because the *macro*scopic view is the one we ordinarily have. I adopt and adapt Dennett's Cartesian terminology in order to address how and why consciousness could seem like a Cartesian Theater phenomenon from the ordinary point of view—and perhaps more importantly, how and why it could seem or feel to us (to itself) like anything at all.

Like a literal camcorder, a brain's Cartesian Camcorder, by this proposal, is an actual physical, mechanical entity, though its resemblance to a literal camcorder is very loose. For instance, as just noted, there may be a number of different constituent recordings that are continually revised with reference to one another. There may be separate modules that coordinate different aspects of memory—short-term and long-term, visual and tactile, for instance. But such implementation details need not concern us for the purposes of this overview.

More importantly, unlike a literal camcorder, the Cartesian Camcorder does not record or play back raw pixels, sound amplitudes and frequencies, and so on, but rather records events at a much higher level of abstraction. The cognitive system parses sights, sounds, and other perceptions into representations that designate physical objects, persons, and so forth. When you see a flower, your cognitive system represents that event in terms that designate not just the flower, but also that you, a person, with such-and-such identifying attributes and history, have just seen a flower; the system represents myriad aspects of what a flower is, and what a person is, and what it is for a person to see a flower, and so on—and the Cartesian Camcorder's recording uses such representations. Thus, the Cartesian Camcorder makes and plays *smart* recordings rather than verbatim recordings.

The cognitive system represents and can calculate various relations among those terms of representation: it makes explicit many aspects of

how the represented objects or states relate to one another, what affected what else, what depended on what else, what else would have gone differently if some event had been different, what events are desirable, how a different choice of action in the future might lead to a more desirable (or less desirable) outcome, and so on. (A proposed sketch of some of the relation-representing machinery appears below in sec. 2.4.1, with elaboration in chap. 5.)

The cognitive system's knowledge of such interrelatedness facilitates the system's pursuit of goals: the machinery engages in prediction and planning, selecting actions to pursue desired results according to its expectations of what would occur if one action or another were taken. The Cartesian Camcorder's recording uses the terms of representation that the prediction and planning systems use. And those systems engage during the playback, figuring out for example what could have unfolded differently, or what might be done differently in a similar situation next time.

Indeed, our ability to learn, to figure out an answer, by replaying past events and thoughts—rather than just being passively entertained, as though by a sitcom rerun—is presumably what would have prompted the evolution of a record-and-playback faculty in the first place. Our current technology is much better than the brain at making long-lived, high-fidelity verbatim recordings, but much poorer at making smart recordings that the machinery can flexibly and autonomously put to use.

The content of Cartesian Camcorder recordings is thus available to the cognitive system's planning apparatus—roughly, the apparatus that implements what we regard as voluntary activity in the pursuit of various goals. Among our voluntary activities is ordinary speech. And in particular, the Cartesian Camcorder recordings' accessibility to the planning system sometimes lets you say things like "I am now conscious of being here, having this thought, seeing this flower . . .".

In contrast, you cannot similarly report "I am now conscious of such-and-such neuron's depolarization" or "I am conscious of parsing the pixels on my retina, figuring out that they show a flower." You could speak those words, of course, but they would not constitute a truthful report of your conscious experience. Those aspects of your brain states or brain processes are not part of what the Cartesian Camcorder is wired up to record, and cannot be viewed by your planning apparatus; that is, they are not *introspectively accessible*.

By this account, consciousness is a property that is endowed upon a cognitive event—endowed upon a perception, thought, emotion, decision, sensation, and so forth—*retroactively* (albeit often within a fraction of a second) by virtue of the event's smart capture and playback by the Cartesian Camcorder. The seeming intrinsic quality of the self-awareness of a conscious experience arises in part because anytime you examine a recorded mental event to see if you're conscious of it, you find that you are. But that's simply because the very examination—that is, the smart recording and playback of the event—itself ensures that the event is conscious.

In this regard, consciousness is like the light in the refrigerator. Whenever you check to see if the light is turned on, you find that it is, creating the illusion that it's always already turned on, when in fact it is the very process of checking that then turns it on. Similarly, you cannot examine a mental event without thereby finding it to be conscious. That consistent detection helps create the illusion that consciousness is always already present in the very event being examined;[5] thus, consciousness is seemingly an intrinsic property of that event. Actually, though, it is by virtue of the very examination that the event is conscious.

But in contrast with checking on the light in the refrigerator—an action that causes the light to turn on—the examination of a mental event does not exactly *cause* the event to be conscious. Rather, the examination of a mental event (i.e., the smart recording and playback) is what *constitutes* that event's consciousness.

Bennett Cerf had a children's riddle: what's big and red and eats rocks? The answer: a big, red rock-eater. One way to interpret the joke is that we expect to learn of some strange new being that exhibits the specified properties. Instead, the entity is defined tautologically as anything with those very properties—perhaps just a familiar animal (or person) with an unusual pigment and proclivity.

Similarly, suppose you happen to reflect that you just saw a flower, or that you just stretched your arms, or thought of an item to add to your shopping list. That reflection is not a *manifestation* of a somehow already intrinsically conscious event of seeing, acting, or thinking; the myriad pieces of knowledge that that reflection entails (understanding that you, a

5. Minsky (1986) calls this the *immanence illusion*.

person, have just seen a flower, etc.) are not *properties* of an already intrinsically conscious event. Rather, that reflection—that smart recording and playback, by the Cartesian Camcorder, of a given mental event—*is* your consciousness of the event. Had you seen the flower, or stretched your arms, when your attention was focused elsewhere, the recording and playback might not have occurred, in which case those same events would not have been conscious.

To postulate an intrinsic consciousness that merely *manifests* itself in the sort of properties just mentioned is a false *reification*. When we envision an entity behind some observations and construe the observations as manifestations of that entity, we reify (i.e., "make real") the source of the observations (for example, we may, through experimentation, detect the properties of some new subatomic particle and then conceptualize the particle that has those properties). Sometimes, as with subatomic particles, the reification does identify a different kind of entity with the specified properties. In the case of consciousness, though, as with the big-red-rock-eater joke, there is no physically special, new kind of entity lurking behind the properties.

Thus, in the present view, as in Dennett's, there is not a unitary, self-contained Cartesian observer sitting in the "audience," separate from the mental events being observed. But, I would emphasize, there is nonetheless a process of observation—the Cartesian Camcorder's smart recording and playback of an event—that is distinct from the observed event itself, and by virtue of which the event is conscious. To a good first approximation (though not, as Dennett points out, on a microscopic scale), a given event is either in the system or not; that is, it has been recorded and replayed or not; that is, it has become conscious or not.

The Cartesian Camcorder bestows consciousness not because the smart recording and playback somehow *generate* consciousness as some additional, extraphysical phenomenon, but simply because the mechanical process of smart recording and playback *is* the consciousness of the event. That is, the recording and playback is what we turn out to be perceiving when we perceive what we call our consciousness. We need no additional magical spark to explain what we perceive. And thus we need no explanation of why or how the recording and playback becomes conscious, be-

cause there is no additional *becoming* step to explain. Rather, the smart recording and playback is all that's going on there.

The Cartesian Camcorder's playback of an event can itself be recorded and played back. This recursive property also allows for consciousness of consciousness, and so on, without requiring a regress of separate observers—though it does require a sequence of distinct observations. A sequence of recordings of recordings can go on for arbitrarily many cycles. In practice, you probably don't go beyond reflecting on your awareness of your awareness of your awareness—only a few such levels. Still, you recall having just been someone who in turn recalls still-earlier versions of yourself, extending back in a long series. In effect, the camcorder's smart recursive recording does implement a (purely physical) *self* that is both observer and observed—or at least, the camcorder implements a parade of self versions, each of which observes its predecessors.

The recursive nature of the camcorder's recording presumably contributes to blurring the distinction between a perception (or other cognitive event) and its consciousness. The event and its smart playback are indeed the same *kind* of phenomenon, in the specific sense that both are events that are recordable and replayable by the Cartesian Camcorder. And these similar events can occur in rapid succession, within a fraction of a second of one another. The light-in-the-refrigerator confusion further blurs the distinction between what is observed and what does the observing: looking at the observed event, we always find its consciousness, so the consciousness mistakenly seems inherent in the event.

But any purely mechanical, physical event is *not* inherently conscious. That is, thinking X, if thinking X is just a physical event, is not inherently the same as being conscious of thinking X. Consequently, blurring the distinction between a mental event and its consciousness helps create the illusion that consciousness transcends mechanism—that is, certain mental events seemingly must have some magical extra property of intrinsic consciousness.

In reality, the seemingly inherent consciousness of a given mental event is illusory. Consciousness does indeed require something extra, beyond the physically implemented mental event of which you are conscious. However, what is required is itself a particular physical, computational event. A given mental event is distinct from the (at least slightly subsequent) event

of its consciousness, that is, the event (if any) of its smart recording and playback. And that is, in turn, distinct from the consciousness (if present) of the consciousness of the original event—that is, the recording and playback of the recording and playback, and so on.

2.3 The Problematic Arbitrariness of Representation

The foregoing account equates what an agent consciously observes—what it experiences—with what is internally represented, recorded, and replayed via a (metaphorical) Cartesian Camcorder. The recording uses terms of representation that depict the interrelatedness of the things represented, in a manner that is usable by the system to make and carry out plans in pursuit of its goals.

We can imagine a computational system designed along these lines, controlling a robot with sensors and effectors. We can then conceive of the possibility that you and I *are* such systems. That a computational system with a Cartesian Camcorder could come to represent itself as self-aware is enough to account for our so perceiving ourselves if indeed we are just such physical, mechanical, computational systems. What we perceive when we perceive our supposedly extraphysical consciousness is indeed something real, but it is not something extraphysical. Rather, it is the smart recording and playback carried out by a Cartesian Camcorder.

But—even so. Even if a computational system records and replays representations of many of its thoughts and perceptions, there is still something puzzling about the idea that it thereby has a conscious experiences of those events. The puzzle stems from the arbitrariness of what can be said to represent what.

It's possible to construe representations quite arbitrarily, as is nicely illustrated by what is known as a *substitution cipher*. In the cipher, we encrypt a message by providing a *key*—just some series of numbers—and aligning each next letter in the message with each next number in the key.

```
 I  T  H  I  N  K  T  H  E  R  E  F  O  R  E  I  A  M

16 13  0  7 17 25 16 15  5  5 13 23  2  8 24  1  8 20
```

Then we generate the encrypted message by taking each letter and moving forward in the alphabet by the specified number of steps (returning to *A*

after passing *Z*). For example, sixteen steps after *I* comes *Y*, so the *I* paired with *16* at the beginning of the message becomes a *Y*. Transforming each letter in turn, we get

```
Y  G  H  P  E  J  J  W  J  W  R  C  Q  Z  C  J  I  G
```

(With slightly more effort, we could extend the scheme to include spaces and punctuation marks in the encoding.) If the message to be encoded is longer than the sequence of numbers that constitute the key, we can repeat the sequence again, as many times as needed.

To decrypt the encrypted message, we must know the key that was used to encrypt it. Then we simply apply each number in the key to each next encoded letter, moving backward through the alphabet instead of forward (for instance, we move sixteen back from the initial *Y*, returning to *I*). We can thus recover the original message, provided that we know the key.

But here is where the arbitrariness comes in. Suppose, for example, we are informed that the message

```
I  N  T  H  E  B  E  G  I  N  N  I  N  G  G  O  D  C

R  E  A  T  E  D  T  H  E  H  E  A  V  E  N  A  N  D

T  H  E  E  A  R  T  H
```

is secretly an encryption of another message. The encryption uses the following key, which is the same as the last example's key, but with some more numbers appended at the end of the sequence:

```
16 13  0  7 17 25 16 15  5  5 13 23  2  8 24  1  8 20

24  0  8 21 16  9  0 19 19 13 17 24 14 11  6 18 21  3

14 14  0 13 13  3  5 20
```

If we decrypt the above in-the-beginning message using this key (stepping backward through the alphabet for each letter because we are decrypting now, not encrypting), we recover the following hidden message:

```
S  A  T  A  N  C  O  R  D  I  A  L  L  Y  I  N  V  I

T  E  S  Y  O  U  T  O  L  U  N  C  H  T  H  I  S  A

F  T  E  R  N  O  O  N
```

The joke here, of course, is that any message can be construed as a substitution-cipher encoding of any other message (of the same length), simply by contriving the appropriate key. The contrivance is easy: just

align the corresponding letters of the two messages and for each pair of aligned letters, choose as the corresponding key-number however many alphabet steps are needed to get from the ''unencrypted'' letter to the ''encrypted'' letter. That, of course, is how the above key was devised—the biblical scribes were not covertly playing for the other team, planting hidden communiqués. The encoded lunch-invitation message didn't really reside in the original in-the-beginning message, but rather in the key itself (or more accurately, in the combination of the key and the original message). Nonetheless, the original message, together with the key, does serve to represent the encoded message.

Returning to the subject at hand, the point of the encryption joke is that construing physical events (such as brain activity) as representations (of external things such as flowers, and of other brain events themselves) leaves room for similarly mischievously creative interpretations. (Putnam [1988] made this point.)

There is a natural, obvious way to interpret brain activity (or electrical activity in digital computers) as representations:

- Signals that are produced by sensory inputs may be regarded as representations of the external objects that give rise to the signals—particularly when the system responds to the signal in a way that makes sense given that representation (for example, by ducking when the retina shows an image of an incoming stone), and when the sensory apparatus produces the signals it does *because* those signals help promote sensible reactions (see the discussion below of Dennett's *intentional stance*).
- And some cognitive or computational constructs can be construed as performing inferences—for example, *If A and B, then C*. We might identify a piece of circuitry that, when representations of *A* and *B* are active, makes active a representation of *C*, thus justifying our construal of that circuitry as representing the stated inference. (Again, the construal is particularly apt if the system responds sensibly to the inference, and is so constructed *because* its response is sensible.)

Using such interpretations, we could, for example, look inside a computer (or, in principle, a human brain) that is playing chess and identify the electronic (or neural) events that represent particular contemplated moves and board positions, and so on.

But alternatively, we can also construct contrived, unnatural construals of physical states as representations—construals where, as a joke, what is supposedly represented is built into the very scheme of interpretation, as with the lunch invitation above.

For example, we could pick up a random rock and construe it as playing a game of chess. We'd already need to have a detailed description of the series of internal states that a real chess-playing computer goes through in the course of a particular game. Then, we'd just point to as many atoms in the rock as we need and we'd stipulate, for each atom at each moment, that its state at that moment represents a particular constituent state of the chess-playing computer. That is, we'd build a translation table with entries like the following:

If the rock's atoms numbered 458,620,198,259,728 through 458,620,198,570,954 at time *t* are in such-and-such state (the state they were in fact in at *t*), that represents the chess-playing computer's transistor number 11,252,664,293 being in thus-and-such state at *t* (the state it was in fact in at *t*).

Of course, there'd be no uniformity to our interpretation scheme—the same state, exhibited by different atoms, would have an entirely different "meaning" in each case. Even the same state of the same atom would "mean" entirely different things at different times (just as different occurrences of the letter *A* in the faux encoded message are made to stand for different letters, as needed to produce the desired interpretation). And of course, we have no hope of writing down every entry of the mapping table—it's just too huge. Nonetheless, we can speak of the interpretation scheme that the hypothetical table implements.

This joke interpretation of a rock as a chess player has no practical utility:

- Our regarding the rock that way would not help us predict what the rock will do next (whereas regarding a true chess-playing computer as such—especially if we also can estimate its skill level—does indeed give us some idea what it will do).
- Similarly, the joke interpretation also tells us nothing about what the rock would have done differently had the representation of one of the chess moves been different—e.g., what next if the knight had moved instead of the rook? Unlike a real chess computer's representations (under a well-motivated representation scheme), the rock's representations (under the

joke scheme) would not then have adapted, would not have corresponded to a sensible response to the alternative chess move. (This so-called *subjunctive* or *counterfactual* property—addressing what else would be different from how things actually are if some particular hypothetical difference obtained—is examined in detail in chap. 5 below.)

In the same way, the contrived key for the lunch-invitation message does not, if repeated cyclically and applied to a continuation of the original message, yield a continuation of the lunch-invitation message. On the contrary, it yields only gibberish. The contrived key works only for exactly the portion of the message for which it was contrived—it does not project further to any other part of the message.

According to Dennett's (1987) notion of the *intentional stance*, we construe an entity as having intentions, beliefs, perceptions, and so forth if that construal helps us model the entity's behavior. And we construe some aspects of the entity as representations if those aspects' correspondence (under some scheme of interpretation) to their putative referents contributes to the entity's behaving as the intentional-stance construal would have us anticipate. Arguably, a complex enough intentional-stance-supporting correspondence occurs only if the correspondence was designed or selected for *because* the correspondence usefully supports the behavior that is understood via the intentional stance (for example, the neural signals from your retina correspond to visible objects, and the neural circuitry evolved as it did because that correspondence helps support your everyday interactions with the objects you see—e.g., what an observer would construe as your reaching to grasp a pen).

The joke interpretation of the rock as a chess machine clearly flunks Dennett's intentional-stance test—the rock's chess-playing properties (under the contrived representation scheme) did not come about in part *because* those properties then let the rock represent a chess player under the contrived scheme. (On the contrary, of course, it is the particular representation scheme itself that was devised *because* it lets the rock represent a chess player.) And the contrived scheme does not extrapolate to explaining or predicting anything about the rock, just as a contrived encryption key does not extrapolate to a longer passage of text than the key was devised for.

Here, though, is a problem regarding the proposal a few pages back as to what consciousness turns out to be (namely, the smart recording and play-

back of some mental events by a Cartesian Camcorder). The problem is that we could likewise contrive a joke interpretation scheme according to which a rock is conscious. As with the joke chess-machine interpretation, we could (in principle) devise this scheme by taking a conscious entity—say, me—and recording all the states in its brain over a period of several minutes. We then map some portion of the rock onto some part of the brain. And we contrive a mapping function that, at each next moment, just asserts by fiat that the state of a given portion of the rock (the specific placement of individual atoms there, say) represents the next recorded state of the corresponding brain portion's state. Under this joke interpretation scheme, the rock undergoes the same sequence of conscious (and unconscious) thoughts over then next few minutes as I did during the few recorded minutes. Or, we could in principle create a different mapping table that attributes to the rock a series of thoughts and feelings all its own.

Of course, the intentional-stance test easily disqualifies such joke interpretations from being taken seriously (just as with the joke chess-machine interpretation). Still, there is a reason this joke interpretation poses a problem. The intentional stance only tells us what representation (if any) an *external observer* has reason to ascribe to an object of interest. And as noted above, a certain practicality follows from an intentional-stance-supported interpretation: it lets us predict that the correspondence in question would or will continue to exist under a reasonable range of circumstances. In contrast, it is of no more practical value to contrivedly ascribe a consciousness-implementing set of representations to a rock than it is to contrivedly ascribe a chess-implementing set of representations. In that sense, a rock is clearly no more engaged in conscious experience than it is engaged in chess playing.

But if we are asking whether an object—be it a person or a computer or a rock—is conscious, we are asking (at least in large measure) about the object's *own* point of view, about how (or whether at all) the object feels to *itself*, regardless of any external observer's perspective or the practical merits thereof. If, as proposed above, consciousness is just a particular kind of representational process, then why isn't it the case that at least as far as the rock itself is concerned, the rock possesses the same stream of consciousness as I do, by virtue of the (albeit impractical) joke interpretation?

Animists—who attribute life or consciousness to all objects, even rocks—might happily assent to the joke interpretation of a rock. But animism

differs from the joke-representation viewpoint under discussion. Animism is a form of dualism, proposing that consciousness is indeed something extra, some kind of ghostlike spark, rather than a mere physical, mechanical process. Animism simply ascribes this ghostlike extra attribute very generously, to all objects, rather than more parsimoniously to people or other organisms.

Taking a step back, we can ask: why not just accept the consciousness of a rock, under the joke interpretation of what the rock's states represent? What's the problem? Remember, we're not ascribing any extraphysical spark to the rock (or to anyone) by virtue of deeming it conscious, on the representational account. The rock is still just a structure of inanimate atoms, juxtaposable with an arbitrary, contrived decoding that regards the rock as making certain representations and self-representations. So what? Why should we balk at that state of affairs? It's not as though the juxtaposability somehow confers some special worthiness upon the rock, some reason we should care about it, some entitlement by the rock to have its survival or its desires (as assigned to it by the joke interpretation) respected by others.

Ah, but then again, you and I too, according to the nondualist theory advocated here, are just structures of inanimate atoms, juxtaposable with a decoding that ascribes to us the kind of representation and self-representation that constitutes consciousness. Why then should anyone care about *us*? That is, why should *our* survival or desires (as assigned to us by the practical, nonjoke interpretation of what our brain states represent) be respected by anyone, any more than would the existence or ''interests'' of a rock or any other inanimate object? Why should we humans care about each other, or even care about our respective selves, any more than we care about a rock under a joke interpretation? Why should the *kind of interpretation* that ascribes consciousness determine the worthiness of respect? As just discussed, the kind of interpretation—joke versus nonjoke—does have practical consequences for an external observer. But regardless of external practicality, doesn't the mere fact that an entity *feels conscious to itself* confer upon it a certain deservedness?

If (as argued here) dualism is false, and if (as also argued) what we are perceiving when we perceive what we ostensively label as our consciousness turns out to be a collection of physical events that form particular kinds of representations and self-representations—smart recordings and playbacks

by a Cartesian Camcorder—then we seemingly need to clarify whether consciousness is just a matter of exhibiting such representations under *any* scheme of interpretation, or whether it somehow requires a *practical* scheme of interpretation, one that is defensible by something like Dennett's intentional-stance criterion.

This very question, however, should alert us to the likelihood of an inadvertent semantic sleight of hand, as discussed in section 1.2.2 above. You and I exhibit consciousness according to either a definition of consciousness that requires a practical scheme of interpretation, or (of course) according to a definition that is indifferent to whether the scheme is practical. Either such definition is a refinement of the ostensive labeling (where we "point to" certain things, and call those things our consciousness). After all, we are pointing both to something of the more general description, and also to something of the more specific description, when we point to instances of our own consciousness.

But a rock (unlike you and me) exhibits consciousness only by the second definition, the one that is indifferent to practicality. And another accurate description of our cognitive system—even more restrictive than the one requiring a nonjoke interpretation scheme—is having the right sort of representations and self-representations under a nonjoke scheme *and* implementing those biochemically, with neurons and carbon atoms, rather than say electronically, with transistors and silicon atoms. That, too, is what we point to when we ostensively identify our own consciousness. Searle (1980) argues that the more general, computational criterion is insufficient for true consciousness; section 7.3.4 below replies to Searle's position.

Much seems to turn here on just which of these definitions of consciousness we select. If the indifferent-to-practicality definition of consciousness seemingly implies that a rock deserves to be cared about as much as you and I do (or contrapositively, that you and I deserve no more regard than does a rock), then an additional definition of *consciousness* is being smuggled in unnoticed—a definition that regards consciousness as implying a deservedness of being cared about or respected. But whether an entity has such an entitlement is not a matter that can be established by fiat, by decreeing a word's definition one way or another, and then implicitly attaching the word to a second definition, smuggled in by sleight of hand. Rather, what we need here is a *theory* explaining what property (if any) does

entitle some entities to a respect for their existence and their interests—and why. In other words, we need a theory of the foundations of ethics.

I attempt to sketch one such theory in chapter 7, after laying some more groundwork. For now, suffice it to point out that when contemplating physical, computational consciousness—like when first contemplating the roundness of the earth—tendrils of the theory reach out to ensnare some far-flung considerations. Changing our view of the earth's shape requires adjusting some of our physics, recognizing that people on the other side of the earth need not walk with their heads toward the ground, because instead, *up* and *down* turn out to work differently than previously thought. Similarly, a mechanical theory of consciousness may seem at first to imply that you and I deserve no more ethical regard than rocks deserve (either because rocks' importance gets elevated, or ours diminished). But I argue in chapter 7 that we needn't be stood on our heads in that way. Rather, there is a reason that (only) a practically interpreted representational consciousness—which we exhibit, but rocks do not (but suitably programmed computers would, contrary to Searle)—is a sufficient foundation for moral entitlement, contrary to what we may have thought according to older theories that require a ghostlike essence as a precondition for deserving ethical regard.

2.4 Origins of Purpose and Value

Before tackling the difficult question of ethics—the question of why (especially if we are just physical objects) we have a rational obligation to respect one another's existence and interests—it behooves us to consider why a person even cares about that person's *own* interests—or what doing so even consists of.

Consciousness is not disinterested. It is not just a neutral, indifferent representation of self and world, devoid of preference. On the contrary, there are things we care about—sometimes intensely—on a variety of levels, from purely physical sensations and visceral emotions to much more abstract goals.

Value-laden physical sensations provide the most straightforward examples. Introspectively, the sensation of burning your finger is intrinsically undesirable. The sensation of eating food is intrinsically desirable (at least

when you're hungry). You tend to avoid burning your finger—given a choice, you don't knowingly let it happen, except perhaps as a side effect of achieving some other goal that you care enough about. Conversely, the sensation of eating when hungry is something that, because it is pleasurable, you do choose to have happen—again, unless some other consideration that is more important deters you.

But what is it about the sensation of getting burnt that is undesirable? What is it about the taste of food that is desirable? The questions sound a bit odd. Some sensations just seem *inherently* desirable—it's part of their very nature.

In contrast, if I ask you what you like about owning a coat, you can explain its purpose and benefit—that it serves to keep you warm and dry, that it looks attractive, and so on. And indeed, sensations like taste or pain do have a purpose or benefit—they serve in part to promote our health and safety by encouraging salutary behavior (more or less; our taste buds may not encourage the most healthful forms of eating under modern conditions, but at least eating at all is much better than not doing so). Still, even without those indirect benefits—even if we had the power to heal burns quickly and fully, or could survive without eating—the sensations themselves would feel somehow desirable or undesirable *just for their own sake*.

But how can that be so, if sensations are just physical, computational, representational phenomena, ''designed'' by evolution to provide us with useful information? How can a physical phenomenon be intrinsically good or bad (as opposed to instrumentally good or bad as a precursor or obstacle to achieving various goals)? Where could intrinsic value come from?

And from the standpoint of our evolutionary ''design,'' why do our sensations not just feel neutral? Why didn't evolution just let burns and so forth be neutral-feeling informative sensations, but still wire us up with behaviors that carefully avoid circumstances that bring about those sensations? Likewise, why not just wire us to pursue the pleasurable sensations, without necessarily making them *feel* pleasurable (whatever that feeling might consist of)?

The answer to the *why* question turns out to bear on the *how* question, so let us begin with the former. To foreshadow that answer, I argue below that your hardwired tendency to pursue (or avoid) pleasant (or unpleasant)

fingerBurning → *withdrawHand*

Figure 2.1
A situation-action rule prescribes an action to take in a given situation.

sensations, combined with your consciousness of those sensations and your consciousness of your tendency to pursue or avoid them, *is* your conscious experience of pleasure or displeasure.[6]

2.4.1 Pursuing Goals: Situation-Action versus Prediction-Value Machinery

From the standpoint of Darwinian evolution, the (metaphorical) purpose of intelligence—more literally, the reason the machinery of intelligence thrives—is to influence an organism to take actions that are beneficial to the propagation of the genes that give rise to that intelligence. Genes that construct such machinery tend to propagate better than genes that don't, other things being equal. Digressing from the question of consciousness for a few pages, here is a quick sketch of some design considerations for machinery that selects beneficial actions. Then, informed by those design considerations, the discussion returns to the matter of the conscious perception of desirable- versus undesirable-feeling sensations and conditions. (The design of machinery for making beneficial choices also figures into the discussion in chaps. 5 through 7 on how there can be genuine choice —as well as a basis for moral responsibility—in a deterministic universe.)

One straightforward scheme to promote beneficial action is to build into an organism's brain a list of various circumstances (each described in terms of sensory inputs, say), paired with the best action to take in each such circumstance. These pairs constitute so-called *situation-action* rules (fig. 2.1).[7] Each such rule designates a (perceptible) situation and an (effectable) action. The organism's cognitive machinery is set up such that if the rule's

6. The supposed intrinsic quality of conscious sensations is what philosophers call *qualia*. Value-laden sensations—those that are seemingly inherently desirable or undesirable—are a special case of qualia. The strategy just foreshadowed—rejecting alleged value-laden (and other) qualia, substituting accounts of our cognitive machinery's reactions to the sensations in question—is likewise advocated, e.g., in Dennett's *Consciousness Explained*. (See also sec. 2.5.2 below for a discussion of value-neutral qualia.)

7. Situation-action rules appear, e.g., in the *production systems* of Newell and Simon (1972).

situation currently obtains, its action is put into effect. (Some embellishments to the mechanism are helpful to deal, for example, with prioritization when more than one rule's situation obtains simultaneously. But such details are unimportant to the present overview.)

If the rules are set up cleverly, the organism need not figure out for itself what action leads to what outcome. Indeed, the situation-action rules do not even designate any expected outcome. The rules represent only the situations and their associated actions.

Nonetheless, situation-action rules can have the effect of promoting complexly organized activity in which each action's outcome is crucial to deciding what action to take next. To give a simple example that hints at the achievable complexity, consider the rules in figure 2.2. These rules might bring about a series of actions culminating in the swallowing of food. The outcome of each rule's action serves to trigger another rule by bringing about the other rule's specified situation. But the rules do not designate their respective expected outcomes—the outcomes just happen, and the rules are rigged such that each one's occurrence (usually) achieves another rule's designated situation, promoting that next rule's action. Sometimes a given rule's action has several possible mutually exclusive outcomes. The chain of activity just continues in sequence with whichever other rule's situation (if any) matches the actual outcome of the previous action.

Thus, a collection of situation-action rules may systematically promote a series of actions that converge on a particular state (such as ingesting food). If a machine has been designed to converge to certain states—literally

seeFoodLowerRight → *placeHandAtLowerRight*

seeFoodLowerLeft → *placeHandAtLowerLeft*

touchingFood → *graspFood*

graspingFood → *placeHandAtMouth*

foodAtMouth → *bite*

solidFoodInMouth → *chew*

liquifiedFoodInMouth → *swallow*

Figure 2.2
Executing these rules brings food to the stomach.

designed, if the machine is an artifact, or metaphorically designed by evolution, in the case of an organism—we can say that the converged-to states are *implicit goals*. Such machines are *teleological*, that is, goal oriented.

It's not just that a teleological system behaves in a way that converges systematically to a certain state; even a simple damped pendulum does that—whatever angular position and velocity the pendulum starts with, it eventually (after a long trajectory that is different for each distinct initial configuration) stops moving and points straight downward. Unlike a pendulum, a teleological system not only behaves in a way that converges systematically to a certain state—it also behaves that way in part *because* it so converges. That is, the machinery was sculpted by natural selection (or by an engineer's design process) *in order to* get to that target state.

An eye or a camera, for instance, has an image-focusing apparatus *because* that apparatus focuses images—that's why evolution selected the apparatus, or why an engineer specified it. Even organisms that are too simple to have and use explicit representations of situation-action rules are still teleological, with an implicit goal of survival and reproduction insofar as their behavior and metabolism are such as to self-perpetuate, and are that way *because* being that way promotes self-perpetuation.

But alternatively, a machine or organism can be teleological in a more direct, more sophisticated manner. In contrast with a situation-action machine, a *prediction-value* machine uses different units of representation—ones that make the pursuit of goals *explicit*. I have argued in *Made-Up Minds: A Constructivist Approach to Artificial Intelligence* (1991a) that our cognitive machinery (and that of other intelligent animals) is likely of the prediction-value sort. I now briefly describe a prediction-value machine, explaining how its goal pursuit—its ability to achieve desired states—is superior to that of a situation-action machine.

A prediction-value machine is more elaborate than a situation-action machine. In place of two-part situation-action rules, a prediction-value machine uses three-part structures that I've called *schemas*, named after the psychologist Jean Piaget's (1952) term for units of cognitive organization. A schema (fig. 2.3) has a *context*, an *action*, and a *result*. It asserts that when

<*context conditions*> : <*action*> → <*result conditions*> (<*probability*>)

Figure 2.3
A schema has a context, action, and result, designated by the above notation.

its context conditions obtain (the schema is said to be *applicable* at that moment), taking the specified action would lead to the specified result conditions (but not necessarily to the exclusion of other results). The expectation is quantified by a probability designated by the schema—the probability that the result would indeed obtain if the action were to occur under the specified conditions.

A schema thus resembles an extension of a two-part situation-action rule. Its context and action are respectively like the situation and action of the two-part rule, but the schema also has a result specification to make an expected outcome explicit.

However, there is another important difference, apart from the explicitly designated outcome. A schema does not instruct the machinery to take the specified action whenever the specified context conditions are satisfied. Rather, a schema merely predicts something that would occur if the action were taken, given the context conditions.

To arrive at a choice of action, a prediction-value machine also needs an ascription of *utility* (i.e., a value) to various possible states. Utility is numeric and may be positive, negative, or zero. At each moment, the prediction-value machine identifies the likely results (according to the then-applicable schemas) of the actions in its repertoire, and selects the action whose results are the most valued according to the assigned utilities.

More generally, there can be a *chain* of schemas, in which each schema's result conditions include the next schema's context conditions. Starting with a schema that is currently applicable, the machinery can take the action designated by each schema in turn along the chain. If the expected result obtains, then the next schema in the chain will become applicable, so that its result will follow in turn if its action is then taken, and so on. For instance, the situation-action example above can be recast in terms of schemas that chain together (fig. 2.4).

In *Made-Up Minds*, I argued that a plausible neural implementation of schemas—in which each schema corresponds to a physical structure in the brain—can support the fast identification, at each moment, of currently chained-to states (essentially by broadcasting a search signal through the schemas that implement such chains, starting with any currently applicable schemas and traversing any such chains in parallel). The machinery can then identify the most desirable currently chained-to state and the chain that leads to it.

seeFoodLowerRight : *placeHandAtLowerRight* → *touchingFood*

seeFoodLowerLeft : *placeHandAtLowerLeft* → *touchingFood*

touchingFood : *grasp* → *graspingFood*

graspingFood : *placeHandAtMouth* → *foodAtMouth*

foodAtMouth : *bite* → *solidFoodInMouth*

solidFoodInMouth : *chew* → *liquifiedFoodInMouth*

liquifiedFoodInMouth : *swallow* → *feelSated*

Figure 2.4
These schemas chain together to promote eating, provided that *feelSated* is of positive utility.

As a further elaboration of the prediction-value machinery, chains of schemas can be composed or nested. For example, the action *placeHandAtLowerRight* might itself be implemented by a network of schemas that show how to move the hand to that position by a sequence of incremental displacements. The particular sequence to use at a given moment depends on the hand's starting position. I use the term *composite action* to refer to an action (such as moving the hand to a specific body-relative position) that is defined as the achievement of a particular *goal state* (here, the state of the hand's being in the specified position) and is implemented by schemas that chain to that state.

To *activate* a schema is to carry out its specified action when its context conditions are satisfied. The machinery for composite actions identifies a chain of schemas (if any) starting with one that is currently applicable, leading to the composite action's goal state. When the machinery activates a schema whose action is a composite action, the machinery then executes the appropriate sequence of actions to achieve the goal state (fig. 2.5).

Typically, actions are organized at many such nested levels. To obtain food and shelter, you acquire money. To acquire money, you arrive at your place of work. To get there, you drive your car along an appropriate route. To maintain the appropriate route, you turn right at the next intersection. To turn the car to the right, you rotate the steering wheel clockwise. To rotate the steering wheel, you perform a sequence of grasping and pulling motions with your hands. And so on.

handAtUpperLeft : *moveHandRightward* → *handAtUpperMiddle*

handAtUpperLeft : *moveHandDownward* → *handAtMiddleLeft*

handAtUpperMiddle : *moveHandDownward* → *handAtCenter*

handAtMiddleLeft : *moveHandDownward* → *handAtLowerLeft*

handAtMiddleLeft : *moveHandRightward* → *handAtCenter*

handAtCenter : *moveHandRightward* → *handAtMiddleRight*

handAtMiddleRight : *moveHandDownward* → *handAtLowerRight*

etc.

Figure 2.5
Each composite action is defined by the achievement of a particular goal state (e.g., *handAtLowerRight*). Schemas that converge to the goal state implement the composite action.

Yet another elaboration of the prediction-value machinery's goal pursuit consists of what I call *subactivation*. Here, the idea is that instead of activating some currently applicable schema (that is, instead of carrying out the schema's designated action), the mechanism can "imagine" or simulate the activation. (Subactivation is so named by analogy to subvocalization, where you imagine saying something and you "hear" the words internally.) Subactivating a schema forces its designated result into a simulated-achieved state, distinguishable from its actual achievement. A schema whose context state is simulated-achieved is deemed applicable for purposes of being subactivated in turn (but not for purposes of being activated for real); call it *subapplicable*.

Like the process of following chains of schemas, the process of subactivation is a way to explore the space of transformations, linking together the state changes specified by individual schemas, thereby arriving at longer-range plans than are expressed by an individual schema in isolation. Subactivation has the advantage that it can explore novel combinations of outcomes, whereas chaining stringing together is limited to already-formed pieces of a sequence. Let's say the following three schemas are present:

$P{:}A{\rightarrow}R$

$Q{:}A{\rightarrow}S$

$RS{:}B{\rightarrow}T$

If states *P* and *Q* happen to obtain at the moment, then the first two schemas—that share action *A*—are both applicable. Subactivating either schema results in simulating the occurrence of both *R* and *S*, because an aspect of subactivating a given schema is that the machinery notifies all other currently subapplicable schemas that share the given schema's action, and they too set their specified results to be simulated-achieved. With both *R* and *S* designated as simulated-achieved, the third schema is subapplicable, permitting in turn the simulated achievement of its result *T* via its action *B*. However, neither of the first two schemas chains to the third, because neither designated result includes all of the third schema's context conditions. Hence, subactivation can sometimes anticipate an effect of a sequence of actions (here, *A* followed by *B*) when the chaining process, in contrast, is oblivious to that effect.

The greater flexibility of subactivation comes at a cost:

- As noted above, the chaining process works quickly, propagating signals through many chains of schemas at once (which is possible in part because schemas are presumed to be implemented by distinct physical structures that can perform computations and communications in parallel with one another).
- Subactivation, in contrast, is a serial process, because noticing the availability of a schema for subactivation on the basis of a simulated set of states (such as *R* and *S* above) would be inordinately expensive (given certain plausible assumptions about brain architecture, as discussed in *Made-Up Minds*) if there could be arbitrarily many such simulations happening in parallel—the simulations would have to be kept track of separately, because their respective simulated-achieved result-states could not sensibly be mixed together.

The parallel-versus-serial nature of chaining versus subactivation has ramifications concerning consciousness, as discussed in the next section.

The machinery's schema-building apparatus (the details of which I omit here) can learn from subactivation-simulated events as well as from actual events, leading, for instance, to the formation of the schema $PQ{:}A{\rightarrow}RS$ in the above example. Given that new schema, a subsequent path from *PQ* to *T* can indeed be found ''automatically,'' by the fast, parallel process of chaining, instead of requiring another simulation via subactivation. In

effect, subactivation allows the machinery to perform and learn from thought experiments—which are often faster, safer, and less expensive than real-life experiments.

The prediction-value machinery sketched here is far more complex than the situation-action machinery described above—even apart from composite actions and subactivation. Even if we consider just the machinery to use a three-part schema to select actions in pursuit of preferred expected outcomes, that machinery is already more complicated than the paradigm in which a two-part rule simply specifies what action to take in what situation.

Moreover, any collection of schemas using prediction-value machinery could be compiled into an equivalently behaved collection of rules using situation-action machinery—equivalently behaved in the sense that in any given situation, either system would take the same action. (After all, in any given situation, there is some action that the prediction-value machinery ends up taking, even if its process of selecting that action is complicated. A situation-action system could behave the same way in that situation simply by virtue of having a rule saying to take that action in that situation. And similarly for every other possible situation.) Why then, should evolution have gone to the trouble of building the more complicated prediction-value machinery? What selective pressure could there have been to promote that development?

For sufficiently simple, unlearned behaviors—exhibited by insects, for instance—the evolution of a precanned repertoire of tripartite schemas—rather than precanned bipartite situation-action rules—would indeed be implausible. But for more intelligent organisms whose behavior is largely learned rather than hardwired, the opposite holds. A prediction-value cognitive system can deduce (via chaining and subactivation or the like) what action is most beneficial in a given situation. The deduction can proceed from two kinds of information: what action has what effect, and what effect is most beneficial. It is easier to learn separately what would happen if a given action were taken in a given situation—and then to deduce, from preferences among outcomes, what action you should take—than to somehow try to learn in one step, without that decomposition, what action you should take in a given situation. Situation-action rules suffice

to control insects, but more creative organisms need prediction-value machinery, because schemas are more readily learned than are the appropriate situation-action rules.[8]

As discussed above, even a situation-action machine is teleological, with implicit goals. That is, its behavior converges to goal states partly *because* it converges to those states. But there is an ambiguity in this use of the word *because*: it can refer to what *causes* the machine to behave as it does, or to the machine's explicit *reasons* for that behavior. For a situation-action machine, only in the first sense (cause) does the machine behave as it does because that behavior facilitates the states converged to. The second sense of *because* (namely, explicit reason) does not apply, since the machine does not entertain any reasons for its behavior at all (though its designer might have had such reasons—literally or metaphorically for an artifact or an organism, respectively).

A prediction-value machine, in contrast, does have reasons for its actions. That is, it selects actions in pursuit of valued expected outcomes. Thus, both senses of *because* apply to a prediction-value machine: it behaves as it does explicitly *in order to* bring about the goal states thus facilitated; and (as with a situation-action machine, or even a more primitive organism with no explicitly represented rules at all) its behaving that way is caused in part by the fact that so behaving does tend to facilitate those states.

The details of how our cognitive machinery might learn what result an action would have (under specified circumstances), and how the machinery might build schemas accordingly, using empirical experience and subactivation thought experiments, are beyond the scope of the present discussion (*Made-Up Minds* proposes some such details). Our cognitive machinery also needs a way to synthesize new state-elements (which can appear in new schemas' contexts and results), and it needs a way to define new, abstract, composite actions (the action of achieving such-and-such state, as implemented by schemas that chain to that state), thus augment-

8. Of course, the two are not mutually exclusive; situation-action machinery can coexist usefully with prediction-value machinery. We, for example, still have some simple reflexes, which can be implemented by situation-action rules. And even the tripartite structures of the prediction-value machinery can be compiled into situation-action rules for more-rapid deployment, after the tripartite structures have been learned.

ing whatever built-in state representations and action representations we may be endowed with. After all, many of our concepts involve artifacts or social structures that simply did not exist when our species evolved (and hence could not have helped shape the evolution of our ancestors' brains). Other concepts we entertain involve entities such as atoms or galaxies, which were hidden from our species until recently, and which were thus similarly removed from possible influence on the evolution of our cognition.

In *Made-Up Minds*, I explore how representations of novel concepts might be constructed. The required machinery (which again is far afield of the present discussion) goes beyond defining simple logical combinations or similarity measures among existing concepts. As the machinery augments its conceptual vocabulary by the construction of new state-element representations, it has all the more need for learning machinery that organizes the newly represented states into schemas. So the need for a prediction-value mechanism, rather than a situation-action mechanism, is all the more pronounced in systems that are sophisticated enough to invent new concepts.

The machinery for choosing an action on the basis of the desirability of the expected outcome exhibits two kinds of value: *intrinsic* and *instrumental*. Let us say that a state is intrinsically valued (positively or negatively) by virtue of being assigned a numerical utility by the machinery. A state is of instrumental value (at the moment) if its achievement chains (via schemas) from currently satisfied conditions to something of intrinsic value, so that the instrumentally valued state serves as a prerequisite or subgoal. But a machine that invents for itself many of its state-representation terms would also benefit from a third kind of value, which I call *delegated* value.

Among the states you might value are those that correspond to such modern phenomena as having money in your bank account in case you need to buy something, having fuel in your gas tank in case you have to drive somewhere, having an umbrella in case it rains, and so on. Our cognitive representations of these states are plainly among our constructed rather than innate conceptual vocabulary. Since their very representations are not innate, the values that our machinery attributes to them also could not be innate.

On the other hand, their value is not entirely instrumental in the above sense, either. Instrumental value, in the sense of the chaining or nesting of schemas, involves a sequence of actions that, starting now, leads to a specific valued state. Fuel in your gas tank is indeed of instrumental value, in that sense, when there is a particular trip on which you want to embark now; money in your bank account likewise has instrumental value when there is something in particular you want to purchase now. Typically, though, such states (a full tank or bank account) are valued even in the absence of an immediate goal—even when, at the moment, you have nowhere to go and nothing to buy. The states are valued anyway, because such a goal will likely arise at some indefinite point in the future.

Instrumental value, in the sense of chaining and nesting, is essentially a *tactical* consideration, involving a specifically envisioned sequence of actions starting now and leading to a goal. The longer-term value of having fuel or money, however, is more a *strategic* consideration. We tend to treat strategically valued states *as though* they were intrinsically valuable—that is, we try to achieve those states even when they do not at the moment sit along a specific path to other intrinsically valued states. We benefit from doing so because we thereby tend to have those strategically valued states available in general, and therefore available in particular at those times when the states do become tactically (i.e., instrumentally) useful.

The game of chess nicely illustrates the differences between tactical and strategic considerations:

- The goal is to checkmate the opponent, and to avoid being checkmated oneself. Checkmate thus has intrinsic value. Sometimes, near the end of the game, it is practical to envision an exact sequences of moves that could lead to a win (if I move there, she can respond with this or that move; if this, then I can make one of such-and-such moves; if that, I can make one of so-and-so moves . . .). Working out such sequences is tactical reasoning. The intermediate states along a path to victory thus have instrumental value.
- Usually, though, the permutations of possible moves until the end of the game are far too many to enumerate (even for a digital computer, which vastly outpaces humans at such a task). We thus resort to attributing value to intermediate states such as having your queen on the board, or having your pieces arranged to control the center of the board. Even though you have in mind no specific sequence of moves leading from those intermedi-

ate states to the intrinsically valued goal state of checkmate, you anticipate that achieving the intermediate states will, in general, eventually tend to facilitate checkmate. This is strategic reasoning, and the value attributed to the strategically intermediate states is what I call delegated value.

Delegated value is like instrumental value in that its point is to facilitate other things of (ultimately intrinsic) value (although the facilitation is strategic rather than tactical). But delegated value is like intrinsic value in that a state to which value has been delegated is (by stipulation) itself treated as a goal state for tactical reasoning—as when we envision the exact sequences of moves in a contemplated exchange to ensure, for example, that the exchange preserves your queen.

How, then, might our cognitive machinery be designed to decide when to delegate value to a given state? In particular, how might it delegate value to a state that is not even innately represented, a state whose representation the cognitive machinery has synthesized by itself?

Perhaps surprisingly, the mere fact that a given state is frequently of instrumental value (as having money or having fuel are) is not enough to warrant delegating value to that state. The state of having my right foot raised, for example, is often of instrumental value: it must occur repeatedly whenever I achieve the goal of walking forward, which in turn is of instrumental value to many other frequently arising goals (answering the door, getting a glass of water, going to the grocery store, etc.). But it would be pointless (not to mention funny looking) for me to treat this frequently instrumentally valuable state like an intrinsically valued state, and thus generally keep my right foot raised in case a need to walk arises.

For delegated value to make sense, further conditions must hold:

- The state in question must tend to persist, as must the state's negation.
- The duration of its persistence must be at least on the order of the expected time until the state is next of instrumental value.

Such is the case for the strategically valuable states above, but not for the raised-foot state. Under the further conditions, our cognitive machinery should be (and plausibly is) wired to *delegate* value from the typically instrumentally facilitated states to the persistent state that strategically facilitates them. Delegation assigns value to the strategically facilitating state, depending in part on the value of the facilitated states and the frequency of the facilitation. The prediction-value action-selection machinery then

treats delegated value like intrinsic value, so that states with delegated value tend to be pursued as though for their own sake, even at times when the states do not lie on any specific recognized chain to something else of value—for if you were to wait for such a chain (for instance, if you didn't fill your tank until you had somewhere to drive), it might well be too late.

The strategic advantage of delegating value does come at a cost. Although value delegation is designed to promote the achievement of other, already valued states, the very delegation of value changes the value landscape, somewhat shifting the overall mix regarding what outcomes are preferred. Under some circumstances, the achievement of states with newly delegated value might compete with the achievement of the original set of states. The goal-pursuing machinery is thus reoriented a bit; it is no longer pursuing quite the same set of goals as before.

But such is life—indeed, literally. That is, natural selection gives organisms the metaphorical, implicit goal of survival, of propagation, of replication. But each next step in evolution, even if it is successful in improving fitness, thereby introduces a small departure from what it is that's being replicated—the improvement itself constitutes an inexactness of the replication. There is a compromise: by obtaining only an approximate replication, evolution often obtains a far more robust (near) copy than if the replication had been exact. (In that sense, we might say that evolution implicitly delegates some of the implicit value of making a set of replicas—delegates it to making a larger set of near replicas.) All this occurs without any explicit guidance or intention, of course.

But I propose that something like delegated value *is* explicitly represented in advanced prediction-value cognitive systems. Using delegated value, I propose, improves the machinery's overall pursuit of already designated goals. Although there is some compromise—adding new goals constitutes a realignment of priorities—delegated value's facilitation of already designated goals keeps the machinery reasonably grounded in those goals, rather than allowing arbitrary deviations. Delegated value enables the machinery's goal structure to evolve together with the construction of ever more elaborate representations of the world.

2.4.2 Consciousness of Value

The previous section discussed machinery for pursuing goals by choosing an action on the basis of a preference for its expected outcome. Let us exit

that digression and return to the question of how it can be that the experience of some sensations and conditions feels intrinsically desirable.

The foregoing discussion sketched the design rationale for having prediction-value machinery. Describing that machinery, I have referred to some value designations as *intrinsic* to denote that the machinery is wired to pursue those valued states without requiring an explicitly represented instrumental rationale for doing so (that is, without needing to know that the pursuit of those intrinsically valued states promotes something else of value, such as survival). But neither the label *intrinsic*, nor the existence of machinery that pursues such states without needing any other reason to do so, explains why these states *feel* intrinsically desirable to us, or indeed why they feel like anything at all.

The computational events described above—the identification of potential chains of actions that culminate in a desired state, and the systematic execution of actions to achieve that state—are not inherently conscious. For as previously discussed, no cognitive event of which we are conscious is *inherently* conscious. Rather, the status of consciousness is conferred retroactively by virtue of a cognitive event's being recorded by a Cartesian Camcorder (an additional piece of machinery, beyond the prediction-value system per se), and thus made available for further scrutiny.

But what sort of events are thus recorded? Based on design considerations and on introspective evidence, a plausible speculation is that the recorded events include the activations and subactivations of schemas. Those cognitive events correspond roughly to the planning-and-action system's focus of attention. Activation and subactivation are essentially serial processes—there is but one such event at a time (or at most a small number of them). In contrast, for example, the constituent pieces of the fast parallel search through chained-together schemas do not seem to be recorded and replayed—and hence, we are not conscious of that search. Introspectively, the chain discovered by such a search (a chain that implements the path to a familiar destination in a familiar neighborhood, for example, or a chess move in a familiar board position) just seems to pop into our minds. And from a design standpoint, there would be scant justification for the greater resources required to record all the steps of a massively parallel search, especially if the machinery for watching the playback, and learning from it, has only the resources to monitor one serial sequence at a time (or at most a few). Recording serial events is much more tractable.

The earlier discussion emphasized that not just any kind of recording-and-playback is conscious—it has to be a smart recording, using elements of representation whose relations among one another are comprehended by the cognitive system. Something like the schema machinery may help implement that comprehension. Schemas and their state elements provide *declarative* knowledge of the world's state, and of what affects what, and also *procedural* knowledge of how to attain goals; something like the Cartesian Camcorder provides for the representation of *episodic* knowledge, narrated in terms of the comprehended elements.

As noted above, our cognitive machinery's hardwiring to select actions in pursuit (or avoidance) of what I've called intrinsically valued states and delegatedly valued states does not yet explain why we would consciously experience such states as intrinsically desirable (or undesirable), or experience them as anything at all. Rather, by the present proposal, that conscious experience is explained as is any other conscious experience: by virtue of the right sort of internal recording and playback of the states and their interrelations. And part of what is recorded—if the recording indeed includes the activation and subactivation of schemas—is the series of such events that exhibit the machinery's built-in tendency to act for and plan for the achievement of the positively valued states (or the avoidance of the negatively valued ones).

When you are hungry, for instance, your cognitive machinery is influenced to organize a sequence of actions—walking to the refrigerator, grabbing an apple, taking a bite, and so forth—that culminates in the pleasant sensations of tasting and swallowing foods. Indeed, even before you take these actions, you may contemplate the sequence (via something like subactivation, according to the current proposal). The (actual or contemplated) actions in pursuit of a valued sensation are part of what is recorded and understood (hence what is conscious)—you see yourself in the act of pursuing the goal by virtue of seeing your contemplation or performance of the steps of that act. Similarly with the avoidance (and contemplated avoidance) of negatively valued sensations: other things being equal, your machinery selects actions to prevent such a sensation's occurrence, and those actions' very contemplation elicits an anticipation by the machinery that it would select such actions.

The pursuit (or avoidance) of states of intrinsic or delegated value is intrinsic as far as the machinery's design is concerned—that pursuit (or avoidance) is simply what the machinery is wired to do in response to the specified utility values (or rather, it is wired to do so, other things being equal—that is, unless stronger competing goals intervene). And the machinery can be aware of its own intrinsic tendency to pursue or avoid positively or negatively valued states, because the Cartesian Camcorder records the patterns of behavior—and patterns of (subactivated) contemplation—whereby the machinery exhibits its tendency to pursue or avoid those states. From the recording's point of view, states of intrinsic or delegated value appear to be states that by their very nature are desirable or undesirable.

Thus, the introspective impression that some states are intrinsically desirable or undesirable is another big-red-rock-eater phenomenon, a false reification. Whenever you have an opportunity to pursue or avoid a valued state (or whenever you contemplate having such an opportunity), your planning-and-acting machinery engages accordingly to induce a tendency toward the pursuit or avoidance. That always-accompanying reaction creates an illusion; the illusion is that something about the state itself (the state of having a pleasant or unpleasant sensation) somehow just inherently *deserves* pursuit or avoidance; and seemingly, something somehow just inclines you to respond appropriately to this inherent deservedness by your pursuing or avoiding the state in question: you want to eat chocolate *because* it just tastes *good*; you want to avoid stubbing your toe because that just feels *bad*.

But a merely mechanical state could not have the property of being intrinsically desirable or undesirable; inherently good or bad sensations, therefore, would be irreconcilable with the idea of a fully mechanical mind. Actually, though, it is your machinery's very response to a state's utility designation—the machinery's very tendency to systematically pursue or avoid the state—that implements and constitutes a valued state's seemingly inherent deservedness of being pursued or avoided. Roughly speaking, it's not that you avoid pain (other things being equal) in part because pain is inherently bad; rather, your machinery's systematic tendency to avoid pain (other things being equal) is what *constitutes* its being bad. That systematic tendency is what you're really observing when you

contemplate a pain and observe that it is "undesirable," that it is something you want to avoid.

The systematic tendency I refer to includes, crucially, the tendency to *plan* to achieve positively valued states (and then to carry out the plan), or to plan the avoidance of negatively valued states. In contrast, for example, sneezing is an insistent response to certain stimuli; yet despite the strength of the urge—sneezing can be very hard to suppress—we do not regard the sensation of sneezing as strongly pleasurable (nor the incipient-sneeze tingle, subsequently extinguished by the sneeze, as strongly unpleasant). The difference, I propose, is that nothing in our machinery inclines us to plan our way into situations that make us sneeze (and nothing strongly inclines us to plan the avoidance of an occasional incipient sneeze) for the sake of achieving the sneeze (or avoiding the incipient sneeze); the machinery just isn't wired up to treat sneezes that way (nor should it be). The sensations we deem *pleasurable* or *painful* are those that incline us to plan our way to them or away from them, other things being equal.

Section 2.4.1 above distinguished two senses in which a system's behavior might converge to a given state *because* the behavior so converges. In the stronger sense, applicable only to prediction-value machines, the machinery converges to the state in question by means of the machinery's own analysis showing that the behavior leads to the state in question, a state of positive utility. In the weaker sense—again applicable to prediction-value machines, but also to situation-action machines and generally to evolved organisms or designed artifacts—the behavior's convergence to the given state is part of what causes the behavior to occur in the first place (part of what causes evolution, or a designer, to sculpt the machinery that mediates the behavior); the system itself does not necessarily act on the basis of any representation of what state its behavior leads to.

A similar dichotomy arises with the question *Why should I care about X?* I have an explicit *reason* to care about X insofar as X has instrumental or delegated value—that is, insofar as its pursuit is motivated by the pursuit of something else of value. But instrumental and delegated value pass the buck to other things of value, and the deferral has to ground out somewhere. That somewhere is intrinsic value. For states of intrinsic value, we need not have an explicit reason motivating our pursuit of those states.

Rather, we like them ''just because''; that is, our machinery is just wired up to pursue those states.[9]

2.5 Some Contrasts

Without attempting an exhaustive (or even representative) survey, I would like to conclude this chapter with the mention of two interesting alternative views—aspects of Noam Chomsky's and Roger Penrose's stances on the mind–body problem—and a brief discussion of some contrasting interpretations of *qualia*.

2.5.1 Chomsky and the Missing Body

The eminent linguist Noam Chomsky takes an unusual position concerning the dualist mind–body problem (see Antony and Hornstein 2003). In Chomsky's view, Isaac Newton obviated the mind–body problem in that Newton ''exorcised the machine, leaving the ghost intact.'' Chomsky refers, for example, to the instantaneous action at a distance by which the force of gravity works, according to Newtonian mechanics. This partial remote control did indeed undermine an intuitive conception of machinery, a conception in which influence is always *local*: things have an immediate effect only on nearby things (which can in turn influence other nearby things, eventually spanning great distances). Thus, in Chomsky's view, it is the dualist conception of body, not of mind, that physics has eliminated.

As it turns out, though, Einstein's relativity later rescinded Newtonian nonlocality—the influence of gravity (and of everything else) does turn out to propagate nearby at finite speed. Likewise, quantum mechanics was first thought to have reintroduced action at a distance, but contemporary interpretations have again restored locality (see chap. 4 below for a detailed discussion).

9. Evolution, of course, has rigged our intrinsically valued states (liking food, disliking burns, etc.) to promote survival and propagation, and has done so because (in the weaker sense of cause, rather than the stronger sense of explicit reason) the rigging promotes survival. Thus, it is no mere coincidence that my eating food (rather than starving) helps me survive, even if I eat just because I like it, just because doing so is what I'm wired to try to do, without my having any explicit reason for it, and without having any explicit goal for which eating is a subgoal.

Moreover, even if physics did turn out to be nonlocal, it could still be mechanical in the sense that the universe ultimately comprises a collection of most-elementary objects (particles, strings, whatever); and that for each instance of such an object, its next state (or its gradual transformation, if physics turns out to be nondiscrete) is still a straightforwardly expressible function of its own prior state and the prior state of other elementary objects (albeit not just nearby ones, if physics is nonlocal). A world that is mechanical in that sense is still a world in which, given the past state of all elementary objects, no degrees of freedom remain for a ghostlike consciousness to have any influence. Thus, even a nonlocally mechanical universe can pose the usual mind–body problem.

And indeed, despite his professed dissolution of the mind–body problem, Chomsky suggests that we may face certain *permanent mysteries*—including, in particular, how to reconcile human choice with the physical principles that constrain us—that are suspiciously reminiscent of mind–body problems. Section 6.2.3 below revisits that particular mystery after looking in depth at the nature of choice.

2.5.2 Qualia and Gensyms

Another Chomskian mystery-candidate concerns *qualia*, or supposed essences of conscious experiences. The problem can be crystallized in the questions: do I experience the color red the same way you do? Is it possible that my inner experience of seeing red actually matches your experience of seeing green, and vice versa? How would we ever know? Even if we found that your neural circuitry detects red just as mine does, and treats it similarly in subsequent computations, couldn't your red-green *sensations* be swapped, compared to mine? You seemingly have *privileged access* to your own sensations: only you know just what they feel like; an external observer of your neurons could not, even in principle, find out.

Or to put it another way, Frank Jackson (1982) asked: if you had never seen colors, but you had a complete scientific understanding of the physics and neurophysiology of color perception, would you not still be ignorant of what it would directly feel like to see red or green? Therefore, would you not learn something new about those sensations if you were to see colors for the first time? And does that novelty not show that, say, the conscious experience of red involves something more than the scientifically knowable mechanical processes that take place in our brains?

Dennett (1991), citing Churchland (1985, 1990), replies that a literally complete mechanistic understanding of the experience of red would allow you to anticipate the thoughts and feelings you would be disposed to have in reaction to seeing red, subtly different from what blue, for example, would invoke; thus, you would *not* be learning anything new about red even if you were suddenly to see colors for the first time. Rather, red would look and feel just as you expected (and predictably different from blue). If, for your first color experience, someone showed you a colored card, you would already be able to say whether the card is red. Such a feat seems implausible, Dennett claims, only because the underlying assumption—of an encyclopedic enough mechanistic understanding—is itself implausible, given the sheer bulk of what would need to be understood.

Intuitively, though, it feels as though the essential experiences of perceiving red and green would keep their distinct identities *even if dissociated* from any specific memories and connotations of those colors. And I think that intuition may be correct. But it does not support the view that something more than the brain's computation is involved in those experiences. Rather, to use a computer-programming metaphor, qualia may just be *gensyms*.

In the computer language Lisp, a gensym (short for *generated symbol*) is an object that has no parts or properties, as far as Lisp programs can discern, except for its unique identity—that is, a Lisp program can tell whether or not two variables both have the same gensym as their value (as opposed to two different gensyms, or else other kinds of objects, such as numbers or strings). Arbitrarily complicated structures can refer to gensyms, placing the gensyms in relation to one another (and to other objects). But each gensym is uniquely identifiable even apart from its occurrences in any such structures.

Similarly, I propose, our basic sensations, such as seeing red, may be represented by distinct gensyms (or at least, by gensyms suitably augmented—to support, for instance, comparisons of brightness or hue, in the case of color-gensyms, in addition to pure discrimination of identity). A Lisp program cannot examine whatever internal ID tag distinguishes a given gensym from any other; yet the program can tell whether or not something is indeed the same as that gensym. Similarly, we have no introspective access to whatever internal properties make the *red* gensym recognizably distinct from the *green*; our Cartesian Camcorders are not wired up to monitor or

record those details. Thus, we cannot tell what makes the *red* sensation redlike, even though we know the sensation when we experience it. (Gensyms metaphorically capture an aspect of qualia emphasized by many authors—e.g., in recent work, the transparency spoken of by Metzinger [2003]; and the homogeneity and nondecomposability discussed by Clark [2005].)

Our color-gensyms become embedded in structures that record our visual experiences, that designate associated color-names that we learn in childhood, that trigger various emotional ramifications, and so forth. Such structures elaborate what seeing red means to us; the structures, together with the gensym identity-comparison mechanism, let you ask, for example, if the color of an object you're looking at now (or some color that you're imagining now) matches the color of the majestic sunset you saw last night. Still, the gensym for red makes *seeing red* identifiable in and of itself—just as our intuition says—even apart from all the memories and connotations (though you would not, for example, know the *name* for red if not for those memories).

And conversely, knowing about red-related cognitive structures and the dispositions they engender—even if that knowledge were implausibly detailed and exhaustive—would not necessarily give someone who lacks prior color-experience the slightest clue whether the card now being shown is of the color called red. But of course, nothing nonmechanical is involved here—on the contrary, gensyms are a routine feature of computer-programming languages.

Interpersonal comparisons of qualia (is my red your green?) are simply meaningless, just as it is meaningless to ask whether two separate computers are using the same gensym when there is no machinery to perform or define gensym identity comparisons across the separate computers. For if we were then to construct such machinery, we could arbitrarily choose to make the machinery construe a particular gensym on one computer to be the same as some gensym on another computer, or not.

Ironically, then, our seemingly uniquely privileged access to our own sensations—the access by which we can perceive, but cannot communicate, just what our respective inner experiences feel like—turns out instead to be a private *limitation*. Unlike an external observer of your neurons, you cannot (introspectively) look inside your gensyms to see what lets you recognize each one; each gensym's only introspectively discernable property

is its very recognizability (or in the case of value-laden gensyms, your tendency to pursue or avoid the state in question is also discernable, as discussed above in sec. 2.4.2). We can easily misconstrue this limitation of our introspective access, mistaking it for an inherent property of what we are accessing. This misconstrual is another false reification; it promotes the illusion that you can have a private sensation that—although distinctly recognizable to you—is inherently featureless (except by reference to itself; that is, the red sensation's only feature is its manifest redness). A truly featureless sensation would not be externally analyzable—unlike physical objects and events, which admit of external, objective, scientific scrutiny. But in reality, a gensym's objectively distinguishing features are merely hidden, not absent.[10]

2.5.3 Misinterpreting Gödel's Theorem

In 1931, the mathematician Kurt Gödel proved a startling theorem whose misinterpretation occasionally reasserts itself to confront the proponents of mechanical consciousness. In one such episode, Roger Penrose (1989, 1996) signed on to the view that Gödel's theorem shows that people cannot be mere machines. (Or at least, Penrose argued, people cannot be discrete machines, such as can be simulated by digital computers. Penrose proposed instead that our neurons must rely on the supposedly mysterious aspects of quantum mechanics. More on the latter in chap. 4 below.) A brief digression is in order to show why Gödel's theorem has no such implication.

Gödel devised an elaborate scheme by which arithmetic predicates and propositions—such as *This number is prime*—and inference steps in formal logic—such as inferring *P* from *P*-or-*Q* and not-*Q*—can be encoded as numbers. The existence of a logical proof of a given proposition (starting from the standard axioms of arithmetic) is then equivalent to the existence of a number that encodes the proof. That is, if a proof exists, then so does a corresponding Gödel-code number, and vice versa.

10. Clark (2005) makes this point, but he further argues against the reality of your perception of your own sensations (as opposed to your perception of external objects that give rise to those sensations). I believe, though, that the metaphor of the Cartesian Camcorder, with its recursive recordings, captures a sense in which you (often) do perceive (among other things) your sensations—but not their implementation details.

Suppose the number *g* Gödel-encodes the predicate *G*, defined as follows:

G(x): There does not exist a number that encodes a proof (using the standard-arithmetic axioms) of the proposition that the predicate encoded by x, applied to x itself, is true.

Define the *Gödel proposition P* to equal *G(g)*—that is, *P* applies the predicate *G* to the number *g* that encodes *G* itself. Thus, the Gödel proposition *P* says that there does not exist a number that encodes a standard-arithmetic proof of the Gödel proposition *P* itself. But as just noted, every (correct) proof can be encoded as a number in Gödel's scheme. So if no such number exists, then no such proof exists. Thus, the Gödel proposition *P* effectively asserts that *P* itself has no proof. Clearly, then, *P* must be true, because its falsehood would lead to a contradiction: *P* asserts that there's no proof of *P*, so if that's false then there *is* a proof of *P*. But if there's a proof of *P*, then *P* must be true. Therefore, *P* can't be false, because if it were, it would have to be true as well, which is impossible. *P* can't be false, so it must be true.

What does the Gödel proposition have to do with the nature of the human mind? Supposedly it means the following. We humans can see that *P* must be true (by the informal argument just given), even though there can be no formal standard-arithmetic proof of *P* (since the nonexistence of such a proof is precisely what *P* asserts—and we just saw that *P* must be true). If there *had* been a formal proof of *P*, our insight that *P* must be true would be compatible with the possibility that our minds are mechanical, because it is known how a mechanical system can carry out a formal proof—many computer programs do so—so that could be how *our* machinery knows that *P* is true.

But since there can't be a formal standard-arithmetic proof, our insight (about the necessary truth of *P*) must (according to this argument) have some other basis—a basis that transcends logical inference from the standard axioms of arithmetic. And it doesn't help much to suppose that we use a proof that starts with some additional, nonstandard axioms. As Penrose points out, for any such axioms we use, we can define a new Gödel proposition with regard to proofs that start with *those* axioms, leaving us with the same paradox: we can prove that the new proposition must be true, but a formal inference system starting from our new axioms cannot so prove. For any axiom system that (we might suppose) serves as our

point of departure, we thus turn out to have insight about some necessary truths of that system that cannot be mechanically derived from those axioms.

The principal problem with this argument is that our transcendent human insight that the Gödel proposition must be true is actually mistaken! In fact, part of what Gödel proved is that if we want to, we can postulate the *falsity* of the Gödel proposition *P*, and we will not then have any inconsistency (unless the standard-arithmetic axioms were inconsistent to begin with—which, as Gödel also happened to prove, cannot be formally disproved within that axiom system itself). That is, if arithmetic is consistent to begin with, then we can construct a consistent mathematical model that conforms to all the standard-arithmetic axioms, and *also* conforms to the *negation* of the Gödel proposition. But how is that possible? Didn't we just see that the falsehood of *P* does lead to an inconsistency—namely, that if *P* is false, then there exists a proof that it's true? If *P* is false, how can there be a proof that it's true without inconsistency?

Here's how. It turns out (as Hofstadter explains in *Gödel, Escher, Bach* [1979, pp. 452–456]) that there's a subtle loophole that resolves the paradox. If there's a *finite* mathematical proof of a proposition, then that proposition must indeed be true, because each step in the proof is guaranteed not to introduce a falsehood if the previous steps did not. So when we get to the conclusion, we're still looking at something that's true. But nothing in $\sim P$ asserts that the proof of *P* (or the number that encodes the proof) is finite.[11] If a proof is infinite, you can trace through it for as many steps as you like and still never reach its conclusion. Is the never-reached conclusion of that sort of proof necessarily true? Gödel's theorem (which proves that postulating $\sim P$ cannot introduce an inconsistency) shows that the

11. The standard arithmetic axioms do not rule out a model in which the set of all nonnegative integers consists (in ascending order) of the series 0, a_1, a_2, a_3, ... (continuing forever), and also the series ..., b_{-3}, b_{-2}, b_{-1}, b_0, b_1, b_2, b_3, ... (continuing forever in both directions), with every *b* greater than every *a*. In that case, every *b* is infinite. Penrose (1996, p. 108) replies, in effect: so much the worse for computational theories of mind—we know what sort of numbers we have in mind, even if our formal axiom systems fail to capture the notion. But what reason does Penrose have to think this "knowledge" is entirely consistent, if (see below) we cannot formalize it without contradiction?

conclusion of an infinite proof indeed need not be true, even if our intuition says otherwise.[12]

In the end, the Gödel proposition is just another illustration that human intuitions about infinite sets (sets of successive numbers, successive inferences, or whatever) are incoherent. We've known of that incoherence ever since Bertrand Russell (1903) showed that intuitive set theory is self-contradictory. Consider the set of all sets that do not include themselves. Does that set include itself? If it does, then it can't; but if it doesn't, then it must. What that contradiction shows is that our intuitions about infinite sets do not quite make sense—there cannot be sets that are quite like what our intuitions envision. Our intuitions about the infinite presumably derive largely from our experience with the finite, so it is no wonder that some of those intuitions turn out to be spurious.

Therefore, when mathematicians devise axiom systems to try to formalize our intuitions about infinite systems, they don't want to be *too* faithful to those intuitions, or the formalization would just collapse into self-contradiction. To get a (potentially) consistent formal system, mathematicians have to depart from our intuitions in some respect or other—for example, by allowing the possibility of an infinitely long proof whose conclusion is false even though its premises are true.

If there were indeed some transcendent process, impossible to formalize or mechanize, by which we somehow know mathematical truths, that process would lead to the same sort of self-referential difficulty that attracted Gödel's attention. Consider the sentence *I cannot know that this sentence is true*. Is that sentence true? If it is false, then I *can* know it is true, which is only possible if it is true; its falsehood would thus imply a contradiction, so it cannot be false. Therefore, by that reasoning, I know it is true. But there is an immediate inconsistency: I know it is true, but the sentence itself says I *cannot* know that.

Mathematicians fend off this paradox (and others like it) by arguing that the self-referential sentence involved is meaningless; it cannot even be expressed in our standard, presumably-consistent formal systems. Gödel's

12. We might try to sharpen the paradox by devising an alternative Gödel proposition that refers to finite-length proofs, rather than to proofs in general. But it turns out that the predicate *finite* cannot be defined using just the tools provided by standard arithmetic. Hence, neither that predicate, nor the alternative proposition that uses it, can be Gödel encoded.

genius was to devise a way to express, in a standard formalism, a very similar proposition; but because the existence of a Gödel-encoded proof does not provably ensure the proof's conclusion, the final, contradictory step cannot be taken. Thus, insisting on formalizing troublesome self-referential propositions protects us from the contradictory implications of our informal intuitions.

There is nothing nonmechanical, though, about our self-contradictory "insights" about infinite sets or self-referential propositions. Although we usually try not to, we can easily build a computer system that draws inconsistent inferences. (Decades ago, when computers were mysterious devices that cost millions of dollars and filled huge, locked rooms, popular imagination portrayed computers as inherently logical and incapable of error. Today, thanks largely to ubiquitous, inexpensive PCs programmed by Microsoft, a less reverent stereotype prevails.)

Gödel's theorem was a pivotal mathematical achievement. It showed that there can be propositions about arithmetic whose truth or falsehood is not formally provable from the axioms of arithmetic. But it reveals nothing that is incompatible with the mechanical nature of the human mind.[13] Gödel's theorem demonstrates that some of our intuitions are mistaken, not transcendent.

With Gödel's theorem, as with quantum mechanics (see chap. 4 below), a subtle technical misunderstanding makes the concept of a computational mind in a mechanical world seem paradoxical, contradictory. Indeed, the

13. In a similar attempt, Penrose (1996, pp. 72–76) constructs a formalism that effectively asks you the following question, which you needn't answer, but if you do, you must answer correctly: *Will you do other than answer this question affirmatively?* Saying yes (or no) would be self-contradictory. Hence, you mustn't answer—and therefore, the correct answer is yes, but you cannot say so.

Penrose states all this in terms of a formal model in which the question is posed to a computer program. Penrose equates what the program "understands" with the answer it returns. He thus concludes that a program cannot know the answer; nor can you, if you're a computer-simulable machine. (If you're not, then the above question, as posed to you, can't necessarily be formalized in the manner Penrose describes, so his argument becomes inapplicable.) Yet you do know the answer, if you reason as above. In fact, both you and the program can know the answer; the question's self-referential prediction merely precludes you from (correctly) *reporting* the answer. (Some *other* program, simulating you, could discern and report your knowledge of the answer, without contradiction.)

brunt of this book is to argue that our entire dualist heritage springs from a vortex of such misunderstandings.

2.6 Summary

By the present account, consciousness is a particular physical, mechanical, computational phenomenon. It does not involve a special ghostlike, extraphysical substance that somehow derails some of the particles in our brains from following the otherwise exceptionless mathematical rules that physical particles follow (nor does it involve a ghostlike shadow that somehow coincidentally parallels the physical system without being mechanically constrained itself, or that is somehow generated by the mechanical system).

To be conscious (of one's thoughts, feelings, desires, perceptions, etc.) is to have those thoughts recorded and played back by a metaphorical Cartesian Camcorder—an actual physical system within our brains. The events are recorded using terms of representation that designate their interrelatedness with other events and concepts. The recording is thus of events as they are understood—a smart recording—rather than just a transcription of raw sensory inputs. The very playback of the recorded material is among the sorts of events that can be recorded, making the self-awareness potentially self-referential to an arbitrary depth.

Although consciousness is conferred *retroactively* upon selected mental events by virtue of their recording and playback, the events themselves seem *intrinsically* conscious, in part because the very act of "looking at" a mental event (via the Cartesian Camcorder)—to see if it's conscious, or for any other reason—is what confers consciousness upon it (like the light-in-the-refrigerator phenomenon). Indeed, the very act of "looking" in that way *is* consciousness of the event looked at. Consciousness consists of the very collection of properties that falsely seem instead like the manifestations of a distinct entity (similar to the false reification of the big red rock-eater).

Prediction-value machinery (in contrast with simpler, insectlike situation-action machinery) explicitly pursues goals by selecting among actions on the basis of the desirability of the actions' respective expected outcomes. Some such outcomes feel inherently desirable, but that too is a false reification. It is not their inherent desirability that induces us to pursue them.

Rather, our machinery's wired-in tendency to pursue them is what their seemingly inherent desirability turns out to consist of.

The concepts of prediction-value machinery, Cartesian Camcorders, the light-in-the-refrigerator illusion, and the big-red-rock-eater false reification may help dispel the mystery about how our internally experienced consciousness could be physically implemented—how it could be that what we perceive when we perceive our own thoughts and feelings and perceptions could turn out to be but a particular collection of physical, material states and processes, viewed in a special way by special computational machinery.

Still, there are some nagging problems with this account. The discussion of prediction-value machinery suggests that such machinery is capable of making choices to pursue its designated goals. But if the universe is just a machine, if everything that happens or will happen is mechanically determined, then what room remains for such choices to make any difference? This question comes up again in the next chapter and is addressed at length in chapter 5.

The other basic problem noted above concerns the arbitrariness of representation:

- Mental states can ''naturally'' be construed to represent what they were designed or adapted to represent—roughly, what they detect (in the case of perceptions), what they control (in the case of actions), and the truth conditions they preserve (in the case of inferences). When a system lends itself to such a construal, the construal can usefully predict what the system will do or would do.
- But there are other, arbitrarily contrived, joke interpretations by which even a rock can be construed to be encoding a representation of a chess game or a daydream or a grocery list or a subtle philosophical deliberation. Such interpretation schemes have no practical, predictive value. But practicality for external purposes does not dictate how something feels to itself—its own consciousness.

Construing a rock as conscious via a joke interpretation is paradoxical only insofar as it seems to suggest that we should therefore respect and care about rocks (or, inversely, that there's no reason to respect and care about *us*, as we too are just atoms and molecules that are conscious according to

some interpretation of what our brain states represent—albeit a more natural and practical interpretation). Resolving the paradox requires a theory of what we are obligated to respect or care about, and why—that is, a theory of the foundations of ethics. Chapter 7 addresses this fundamental topic.

But first, the next two chapters shift the spotlight from mind and consciousness per se to some paradoxes about physics in its own right. Even within the realm of physics itself, the concept of a mechanical universe leads to puzzles that need to be resolved. But the particular puzzles set out next do turn out to have interesting ramifications concerning our conscious existence and experience.

3 Going without the Flow: The Frozen Stream of Time

The previous chapter looked at our perceptions of our own minds, addressing the paradox of how inanimate, mechanical matter could implement the thoughts and feelings of which we are conscious. This chapter turns instead to two paradoxes concerning what we perceive of the external world, hence paradoxes of physics itself—paradoxes about how it could be that the universe is specifiable by simple rules, and initial conditions, of the sort that seem to describe it, and still be consistent with what we observe with our own senses.

The first such paradox is that the apparent forward progression of time is conspicuous in our perceptions, yet the physical equations that match the universe have no forward–backward temporal asymmetry, and no progression or flow of time at all. The second such paradox concerns the apparent nondeterminism and nonobjectivity of the world as described by quantum mechanics. This chapter and the next address those two paradoxes in turn. These may not be the only physical paradoxes that can challenge the mechanical-universe paradigm, but they are perhaps the most accessible, well-known examples that do foment skepticism about a mechanistic worldview. Their resolution is at least a good start at combating that skepticism.

3.1 Static Spacetime

Imagine that space is just two-dimensional. And imagine that it is a (huge) finite square, rather than unbounded. So the total state of the universe, at a given moment, can be drawn on a square piece of paper (graph paper in the discrete Fredkin–Wolfram view, or unlined paper in the original Laplace view). The piece of paper shows the state of every particle in the universe.

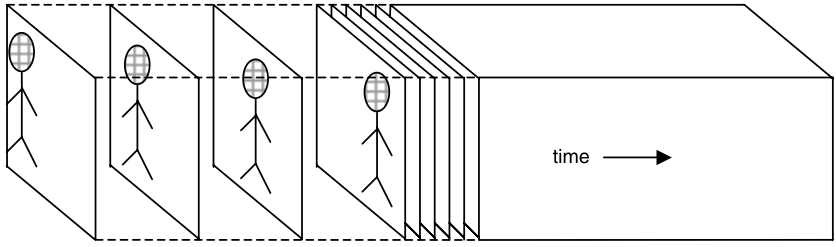


Figure 3.1
We can think of spacetime as a box. Time is one dimension of the box, and all events just sit statically in spacetime. *Change* is just a contrast between two (static) time-slices of spacetime.

By applying patterns that say how the state of each particle changes over time (as a function of other nearby particles), we can generate a stack of pieces of paper, each next page describing the next state of the universe, so that the pieces of paper form a three-dimensional box shape (fig. 3.1). That's a discrete-time model; or, with a continuous model, we still get a box shape, but one from which we can take an infinitesimally thin slice at any point, showing the current state of the universe at that point. Either way, though, we have used two dimensions to represent space, and a third to represent time.

From this boxlike point of view, the laws of physics are just patterns that say how the particles in each cross section are related to the particles in the immediately adjoining cross sections. If the physical laws are deterministic, then any single cross section suffices to establish what the next one must be, and the one after that, and so on, for the entire future of the universe. And to an external observer looking at the two-dimensional universe from outside the three-dimensional box, everything is just sitting statically, unchanging, within the box.

Of course, the successive cross sections of the box might depict a great deal of motion, or other change—as, for example, successive frames of a film depict motion, even though the frames are just sitting there all at once, not changing. An object in motion (relative to whatever reference frame is given by the cross sections) merely occupies different positions within successive cross sections that designate successive times. By contrast, a stationary object (relative to that reference frame) occupies the

same position in successive cross sections. What we think of as change consists of differences between one (static) cross section and another.

The real universe can be thought of similarly,[1] but with three spatial dimensions (or even more, according to string theory) rather than two, plus the extra dimension for time. The three real-world spatial dimensions are not necessarily bounded (like a square or a cube), but rather may extend indefinitely. The resulting four-dimensional ''box'' is impossible to visualize, but it can be understood by analogy to the three-dimensional box for the two-dimensional (plus time) universe. Different cross sections of the four-dimensional box correspond to the universe at different times—but now, each cross section is a frozen three-dimensional expanse extending across the entire universe, rather than a frozen two-dimensional expanse.[2] Just as the completed reel of film does not change, neither does the content of the four-dimensional universe box.

1. According to early interpretations of quantum mechanics, the rules that the particles obey don't specify exactly what each particle does, but rather specify a probability distribution of possibilities. Quantum uncertainty challenges the deterministic, static-box model just discussed. But there is a newer interpretation of quantum mechanics that preserves determinism, albeit in a much larger box—a so-called configuration-space box instead of a physical-space box. Chapter 4 discusses this issue extensively.

2. Readers familiar with special relativity might worry about this picture's designation of time-slices with respect to a seemingly absolute time-axis. How much is undermined if we go beyond modeling Newtonian mechanics? According to special relativity, observers in different inertial reference frames see different orientations for the time axis—an orientation in a Reimann geometry, rather than the familiar Euclidean geometry.

It turns out, though, that although special relativity is conventionally formulated in terms of Reimann geometry, it can equivalently be formulated in Euclidean terms, merely replacing $\boldsymbol{F} = m\boldsymbol{a}$ by the relativistic version $\boldsymbol{F} = m_0(1 - v^2/c^2)^{-1/2}\boldsymbol{a}$. True, we've then implicitly picked, for our representation, one inertial reference frame from among infinitely many equivalent ones (making a correspondingly arbitrary choice about which way to point the time axis). But that's no more worrisome than (in a Newtonian universe) arbitrarily selecting the directions in which to orient the x, y, and z axes. We don't thereby assert uniqueness or absoluteness about the particular choice we've made. Rather, the point is merely that the universe can be expressed in terms of *some* choice (the more the merrier) of axes such that physical laws specify a series (or a continuum) of time-slices with respect to those axes.

The boxlike point of view, with all the past, present, and future sitting statically in spacetime, contrasts sharply with our perception of time. To us, there is always a *present moment* that is somehow more real than the past and the future. The past is just a memory, the future just a potentiality. And crucially, which moment is the present one *changes*, moving steadily forward in time. It is as though the cross sections in the four-dimensional box were indeed frames in a reel of film, and the film is being shown one frame at a time—the present frame.

An imaginary observer external to the whole four-dimensional box might indeed choose to observe the universe one frame at a time—or even to derive the universe, to compute its content, starting with an initial cross section and then applying the laws of physics, one frame at a time. But any frame-by-frame scanning (or derivation) by an external entity—even if it actually occurs—does not correspond to a physical process within the universe itself. The scanning process itself would be entirely external to the universe and hence not in any way detectable within the universe.

Indeed, a hypothetical external observer might even elect to scan the universe *sideways*, looking at successive cross sections along an arbitrary spatial dimension, each such cross section spanning all time. Or the observer might lay out all the cross sections and look at them at once, or scan them backward. No such scan of the spacetime box would make the content of the box different from what it would be if the scan were performed differently, or not at all.

To put it another way, consider two versions of the same universe. In the first version, successive frames are generated (by a process external to the universe) on the fly and quickly discarded, each frame computed from the previous one using the laws of physics. There is only one frame at a time, just as in our intuitive model of the progression of time. In the second version, there is just a static collection of frames—frames with the same content as those of the first universe, depicting the same series of events related to one another by the same laws of physics, but with the frames sitting there all at once, without any serial process of generating or scanning.

Since all depicted events are the same in both versions, it follows in particular that if either version contains beings like ourselves who perceive an apparent flow of time as you and I do, they will do so and say so in both versions of the universe (because their perceptions, and their commentary

on those perceptions, are among the events identically depicted in the two versions). Most importantly, they will do so for the same reason in both versions—namely, because their doing so is somehow part of what their universe's laws of physics ultimately dictate will happen. Even if there is, in one of the two versions, an external scanning or generating process that coincides with the universe's denizens' perception of the progression of time, that perception is not in any way due to that process—even in the version where that process occurs.

For the denizens to propose that the scanning or generating process occurs at all, then, is a superfluous hypothesis, just as the postulated non-negligible influence of gravity on photons was superfluous to explaining mirror asymmetry in section 1.2.3 (and just as postulating extraphysical consciousness is superfluous to explain our impression that such consciousness exists, if the laws of physics are indeed exceptionless, as discussed in sec. 2.1). And the same applies, of course, to us denizens of *this*, the actual universe. There could be, for all we know, some external process that scans or generates our universe, one moment at a time, in accord with our perception of time's progression. But that perception itself is no reason to think so, because even if the external process were real, the perception is caused by something else entirely. (And there is little other reason to posit such an external process, apart from that perception.)

It remains, then, to explain how the mistaken perception of a progression of time arises from principles of physics that do not describe any such process (just as, in the mirror paradox, we needed to explain how it appears that mirrors swap left and right but not top and bottom, when the behavior of mirrors is governed by physical processes that have no such asymmetry). Before addressing that question, it will be helpful to compound the paradox with a second, closely related one, and then solve both together.

The second paradox—even more reminiscent of the mirror problem—has to do with the apparent asymmetry of the *direction* of time. Time not only seems to progress, it seems to do so inexorably in the same direction, toward what we call the future. But the laws of physics that specify the occurrence of all events—events including our perception of time—are actually *symmetric* between past and future, in a sense elaborated in the next section. So we need to explain not only how it is that we perceive a flow of time at all, but also how it is that we perceive a dramatic asymmetry

between past and future—but not, say, between one spatial direction and its opposite.

3.2 Time Symmetry

We remember the past, but not the future. Although there are aspects of the future that we can anticipate, a great many details—the outcome of a particular coin toss, or the weather on a particular day more than a month away—can be known with great reliability afterward, but not beforehand. If we want to get to the near future—say, a minute from now—we need only wait a minute and we're there! But nothing we can do will return us even a few seconds into the past. Perhaps most strikingly, if we watch a movie—of people, vehicles, or animals, or even of many inanimate processes, such as a rainfall or a volcano or an avalanche—we can immediately discern whether the film is being shown forward or backward.

Yet the physical laws of our universe—the laws that say how particles move about and change state—appear to be entirely time symmetric: they treat the forward direction in time as a mirror image of the backward direction. Paradoxically, time-symmetric underlying laws somehow give rise to a sharp asymmetry between past and future with regard to many ordinary phenomena in our experience. As with the mirror paradox in section 1.2.3, this paradox may seem either to counter the claim that the underlying laws are indeed time symmetric, or to call into question the accuracy of our perceptions, or even to challenge the very reducibility of the phenomena we perceive to what is specified by the underlying laws.

But as in section 1.2.3, I argue that there is indeed a way that the asymmetry of the events we observe can arise from symmetric underlying laws. The underlying symmetry is real, but there is also something real about the perceived asymmetry. And it is precisely by reducing the latter to the former—by figuring out how the symmetric laws give rise to the perceived asymmetry—that we can understand how the two are compatible.

Thus, the strategy below is to construct a simple, artificial universe that has time-symmetric laws, and to show that perplexingly time-asymmetric events do somehow arise from those laws. The conclusion, then, is that the asymmetric events of our own universe may be similar, requiring no actual asymmetry of time itself, and no forward flow of time (indeed, no flow at all).

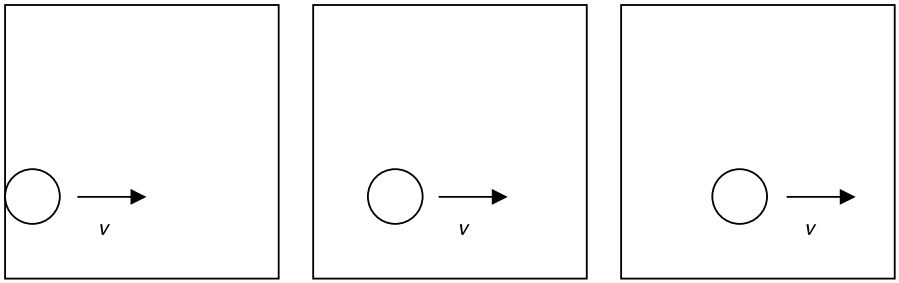


Figure 3.2
In the absence of any intervention, a Newtonian object maintains constant velocity, depicted here by the arrows. (Here and in subsequent diagrams, each square outlines just a small portion of the universe; the square's boundaries are not real.)

To start with a simple, idealized example, consider a disk sliding left to right on a smooth, frictionless surface (think of a perfect air-hockey table), as in figure 3.2. Newtonian mechanics (which describes approximately how the objects around us move and interact at familiar, much-slower-than-light speeds) dictates that the disk continues to slide at the same velocity until some force disturbs it. Applying that law of motion to the present state, each next state gets updated by moving the disk a bit to the right (according to its velocity) and leaving the velocity unchanged.[3] But similarly, we can apply a mirror image of the law of motion to derive the immediately past state from the present state. The mirroring consists of using the *negative* of the disk's velocity to update its position. Hence, tracking backward in time, we see the disk move slightly leftward rather than rightward at each step. Thus, the backward-in-time law of motion is the same as the forward-in-time law, except for the reversal of the velocity's sign.

Laws of motion also specify what happens when objects collide (fig. 3.3). Here, too, the laws preserve time symmetry. For two idealized disks of equal mass that neither deform nor heat up when they collide (such collisions are said to be *elastic*), the rule is simple. Draw a line that is tangent to the two disks at the point where they come in contact. The disks exchange the component of their velocities that is perpendicular to the tangent line and

3. Or rather, that description corresponds to a discrete version of Newtonian mechanics. With continuous physics, there is no distinct "next" state, but rather a smooth, gradual transition that is nonetheless similar for present purposes. For simplicity, I usually speak here in terms of a discrete version.

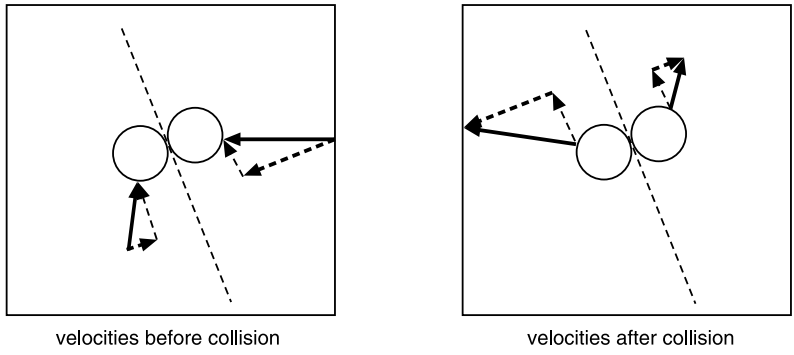


Figure 3.3
The colliding disks exchange the velocity components (the bold dashed arrows) that are perpendicular to the disks' shared tangent. The lighter dashed arrows show the components parallel to the shared tangent, and the solid arrows designate the disks' overall velocities.

retain the component that is parallel, as in figure 3.3. (By the Pythagorean theorem, this rearrangement of perpendicular components of the velocities does not alter the sum of the squares of those velocities. In other words, the sum of the disks' squared velocities is *conserved* in these collisions—unlike, say, the sum of the magnitudes of the velocities, which is not necessarily conserved. Thus, kinetic energy—defined as $\frac{1}{2}mv^2$—is conserved by this simple collision rule. Momentum—defined as mv—is also conserved, because swapping some of the vector components does not change the sum of all the components.)

The motion laws for collisions also exhibit time symmetry. Aside from the mirror imaging—reversing the sign of the velocities[4]—what happens just previously is given by the same rule as specifies what happens just next. If we watch a movie of a collision that obeys the above rule, the movie looks reasonable whether it is running forward or backward—the colliding objects follow the same rule either way, and there is no visible clue about the direction in which it's running (fig. 3.4).

More generally, consider a universe with many such colliding disks (an example of what W. V. O. Quine [1969] has called a *Democritean universe*)

4. If we consider real-world physics instead of the Newtonian simplification, it turns out that some other sign reversals are needed in addition to the ones for velocities (so-called *CPT invariance*). But the principle remains the same.

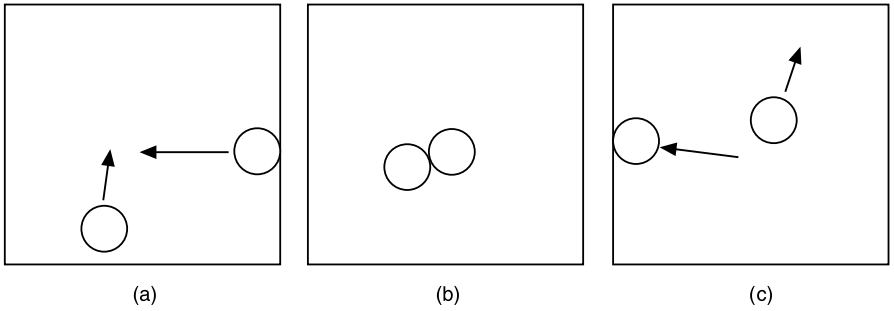


Figure 3.4
This collision sequence follows the same physical laws whether viewed as (a), (b), (c) (with the velocity arrows as shown) or as (c), (b), (a) (with the arrows reversed from what's shown).

with randomly assigned initial positions, speeds, and directions. Assume that the positions are assigned independently (except for preventing two disks from overlapping)[5] and uniformly over the universe's space. Likewise, directions are assigned independently and uniformly over all possibilities, and speeds are assigned independently and uniformly over a given range. In this busier artificial universe, too, no obvious cue distinguishes a forward-running movie from one running backward, as figure 3.5 illustrates.[6] For the same reason, looking at an individual still photo of this artificial universe, there is no way to discern the direction in which any particular disk is moving.

So far, we see what we would expect: time-symmetric laws produce time-symmetric phenomena, which follow the same rules (or rather a mirror image of the same rules) going forward or backward. So far, then, this propagation of symmetry from underlying laws to the phenomena the laws implement is exactly the opposite of what I promised to demonstrate.

5. Let's assume the disks' random positions are assigned sequentially. If a given disk's assignment would have it overlap an already-positioned disk, a new random position is attempted for the given disk, and so on until success.

6. Students of physics will recognize that there is, however, a *nonobvious* distinction between the two directions. As the many objects collide repeatedly, the distribution of kinetic energy will tend toward a Boltzmann distribution. But if the initial distribution of energy were a Boltzmann distribution (instead of a uniform distribution over a range of velocities), then even that subtle asymmetry would be absent.

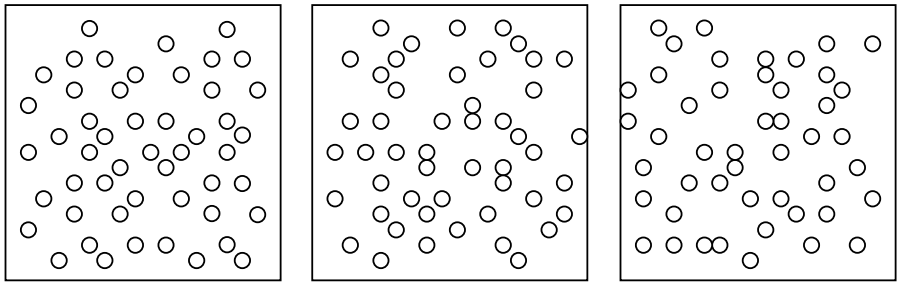


Figure 3.5
Disks collide at random (schematic). Again, sequence (a), (b), (c) obeys the same laws as (c), (b), (a). (Once again, each square outlines just a small sample region of the universe.)

But consider next a slightly more complex universe. It has two kinds of disks: a large number of densely packed, relatively slow-moving miniature disks, and a small number of sparsely distributed, very fast giant disks (of proportionately greater mass). Again, let's posit that the behavior during collisions is as prescribed by Newtonian mechanics (which requires a slightly more complicated collision rule than before). As with the simpler universe, we assign initial positions, speeds, and directions randomly, but with much higher average speeds for the giant disks than for the miniatures (for reasons discussed just below).[7]

This more complex universe's laws of motion, too, are completely time symmetric: looking at a movie of any individual collision between two disks (of the same size or not), we have no way to tell if the movie is running forward or backward; either direction is consistent with the universe's physics. However, looking at the aggregate behavior of a great many disks, we notice a striking asymmetry. As figure 3.6 illustrates, a giant disk momentarily sweeps clean the area it just passed through, leaving a visible "wake" behind it (until the wake fills again with smaller disks passing through at random; the much-greater speed of the giant disks ensures a noticeable delay before the wake fills in). The giant disk's collisions also

7. With regard to avoiding overlaps in the initial assignment, assume that we assign random positions sequentially to the giant disks first, then to the miniature disks.

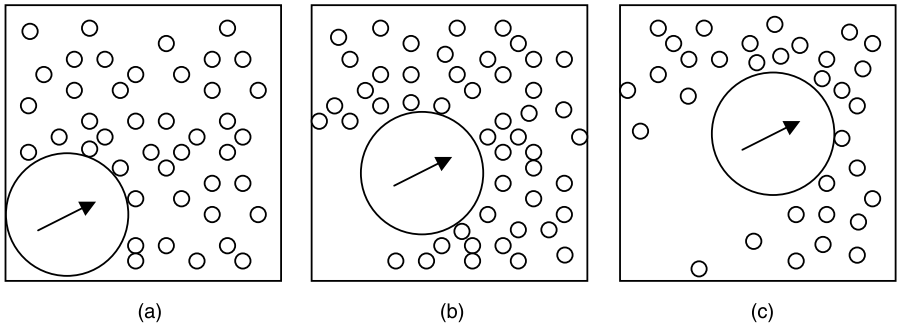


Figure 3.6
A giant disk leaves a "wake" behind it (schematic). The ordering of (a), (b), (c) is apparent.

launch an arc of high-velocity mini-disks in front of it. Those are less obvious, though, at least when we examine an individual snapshot.

Looking at a snapshot, we can tell at a glance which direction a giant disk is moving in: it is the direction away from the wake. Looking at a movie—or at a series of snapshots, as above—we can tell at a glance if the movie is running forward or backward. If it's forward, the wakes trail behind the giant disks; if it's backward, the wakes precede those disks. In the backward movie, it looks as though the mini-disks somehow conspire to form spontaneous anticipatory wakes, moving out of the way of a giant disk before it passes through. But of course, if we look closely at the behavior of each disk and each colliding pair of disks in the backward-running movie, we see that they are just obeying the usual (time-symmetric) rules of motion and collision.

Where, then, does the time asymmetry come from? If each constituent event—each collision, and each object's behavior between collisions—is time symmetric, how can the ensemble of those constituents have such conspicuous asymmetry with respect to time?

In the study of physics, the fields of *thermodynamics* and *statistical mechanics* bear on questions of time asymmetry. These fields examine *entropy*, a measure of how disordered things are, in a specific technical sense.

Consider a deck of playing cards. One possible kind of arrangement is for the cards to be carefully sorted, each of the four suits grouped together,

with the cards going from lowest to highest within each suit, and the suits themselves in any sequence. Such a highly ordered state would be said to have low entropy. In contrast, a deck of well-shuffled cards, exhibiting no apparent order at all, would be said to have high entropy. Between those extremes, a deck that is approximately sorted, but with several cards a little out of place, has fairly low entropy, but not as low as a fully sorted deck. Very few possible orderings of the cards are totally sorted; many more orderings are approximately sorted; and vastly more orderings are devoid of any obvious overall pattern.

The technical definition of entropy tabulates how many orderings there are that fall into a given category of arrangements (categories such as sorted, approximately sorted, or well shuffled). The fewer such orderings there are, the less entropy the category is said to have. Thus, an ordering selected entirely at random is more likely to be in a high-entropy category than in a low-entropy category—simply because the high-entropy category, by definition, has many more members. (It is possible, though astronomically unlikely, for a randomly established ordering of cards to turn out to be sorted, for example.) And if entropy is not already maximal—if a deck of cards is not already thoroughly shuffled, for example—then a gradual reordering (say, by a process by which one card after another switches places with one of its neighbors) will, on average, gradually increase the entropy of the collection (for example, gradually moving from fully sorted to approximately sorted to barely sorted to thoroughly shuffled).

Similar considerations apply to, say, a collection of particles bouncing around in a given space. It turns out there are vastly more configurations with an approximately uniform distribution than configurations with areas of conspicuous concentration or sparseness. For example, in the Newtonian universe just defined, there are vastly more mini-disk arrangements that are roughly uniform than that exhibit sparse ''wakes.'' (And in turn, by an even more enormous factor, arrangements that are approximately uniform, but with some wakes here and there, outnumber arrangements that have all the mini-disks confined together in just a small subset of the space.) Thus, we expect a wake formed by the passage of a giant disk to fill in as the random motions of the mini-disk tend toward arrangements of greater entropy. Likewise, we do not expect wakes to form spontaneously.

The formation of a wake by a passing giant disk is an example of how entropy can decrease locally—but only at the cost of a greater increase else-

where. In this case, the much-greater velocity of the giant disks, compared to that of the mini-disks, is itself a low-entropy condition. After very many wake-forming collisions, the disparity will even out, with the giant disks transferring their energy to the mini-disks until both sizes move about with more comparable velocity (at which time the giant disks will no longer generate conspicuous wakes). The formation of a wake is a temporary lowering of the entropy of the mini-disk arrangement, but it is also an increase—by a larger margin, as it turns out—of the entropy of the giant-versus-mini velocity distribution. The entropy of the overall ensemble thereby steadily increases.

Thus, the observed temporal asymmetry of the formation and dissipation of wakes—evident at a glance when watching a movie of the process—turns out to be an instance of the general principle that entropy increases with time. And at first, this principle seems explicable just by appeal to the statistical observation that the higher-entropy states are (by definition) more numerous than the lower-entropy states, and so are more likely to be selected at random (and hence are likely to result when particles are ''reshuffled'' by random collisions).

But this explanation does not resolve the paradox of how a systematically temporally asymmetric phenomenon can arise from fully symmetric underlying processes. Indeed, the observation that higher-entropy states are more numerous would appear, by the same reasoning, to suggest that running the laws of motion *backward* in time would also lead to higher-entropy states. After all, physical laws in either temporal direction specify random local collisions among the various objects, and the high-entropy arrangements enjoy just as large a majority among possible earlier states as among possible later states. Yet somehow spontaneous ''anticipatory'' wakes are ubiquitous throughout the backward-running movie, but virtually never appear in the forward-running movie.

We may be tempted to reply that the anticipatory wake formation only *seems* spontaneous when we watch the movie backward—watching forward, we see the wake was caused by the action of the giant disk. But that reply just begs the question: why, then, can the motion of the giant disk cause a wake behind it in the forward direction in time, but not backward in time, given the time symmetry of the laws of motion and collision? Just invoking thermodynamics and entropy, and the improbability of randomly selecting a low-entropy state, does not yet explain how a time

asymmetry with regard to improbable states can arise from the underlying symmetric laws and symmetric constituent events.

If the laws themselves are time symmetric, perhaps the asymmetry is a property of the initial conditions. Indeed, there is nowhere else the asymmetry could come from, since the laws and the initial conditions together determine the entirety of the universe's spacetime.

But we've been considering an initial state that is itself time symmetric too. There is simply a uniform initial distribution for the giant disks, and one for the mini-disks as well (but over a smaller range of speeds); any resulting configuration is exactly as probable as its time-reversed mirror image in which all the velocities are reversed. Somehow, temporally asymmetric phenomena arise from symmetric laws *and* symmetric initial conditions.

A conventional explanation for time asymmetry is that the initial state has especially low entropy (by virtue of the giant disks' much-greater velocity, compared to the mini-disks, as noted above). From there, overall entropy is overwhelmingly likely to increase. And indeed, this answer—that the initial state had very low entropy—is critical to explaining why conspicuous time-asymmetric phenomena (such as the wakes) occur at all (as they do not occur in the eventual high-entropy state in which the giant-vs.-mini velocities have evened out).

But there remains a paradox about the time-asymmetry of the whole process. Consider a time-slice of the universe a few moments after the initial state. Entropy has increased some, but we still find a very low-entropy arrangement: the giant disks have not yet transferred much of their energy to the mini-disks. Why, then, do we not see an increase in entropy when we apply physics backward from that low-entropy state, as we see when we apply physics forward? Why do we instead see decreasing entropy (as manifested by seemingly spontaneous wakes) when we watch backward from there for a few moments?

Could it be that the few-moments-old state's entropy is not quite low enough to bar this backward increase, whereas the initial state's slightly lower entropy was just below some critical threshold? No, that can't be right, because the discussion so far hasn't even tried to quantify the initial degree of entropy under consideration (it didn't specify *how much* faster the giant disks are than the mini-disks). We were able to infer, from entirely qualitative considerations, that an initial state of the specified form would

show a universe with wakes forming behind the giant disks as we watch a few moments forward in time, but with spontaneous anticipatory wakes instead as we watch backward again from the few-moments-old state. Therefore, the entropy does not have to compare favorably with any specific threshold to generate the apparent temporal asymmetry.

What property, then, of the few-moments-old state accounts for entropy increasing from there in the forward temporal direction, but decreasing in the backward direction? And again, how can such an asymmetry arise from the time-symmetric initial condition and underlying laws?

There is an important clue to be had by looking more closely at the asymmetry itself. In particular, consider the backward-running movie, in which wakes spontaneously form in advance of the giant disks' passage. Each step along the way in the backward movie, we are applying the same laws of physics as in the forward version (except for the reversed velocities).

Suppose we do so all the way back to the initial state—and then keep going (fig. 3.7). What we see, of course, is just a seemingly forward-running movie from that point onward. That is, as we watch further backward in time, past the initial state, the wakes no longer spontaneously precede the giant disks. Instead, they form behind each giant disk as it pushes the mini-disks aside, as in a forward-running movie, even though we are still moving backward in time (that is, we are still looking at snapshots that are successively further in the negative-designated direction along the time axis).

Here, the origin of the simple universe's time asymmetry becomes apparent:

- The seemingly forward direction in time is the direction such that, if we watch a movie running in that direction, we see a wake that follows each giant disk, rather than spontaneously preceding it. More generally, it is the direction of increasing overall entropy.
- But the seemingly forward direction is not a fixed direction at all. Instead, it is a pair of opposite directions, both pointing away from the initial state.

After all, what we see as we watch the movie backward past the initial state is identical to what we would see watching the movie forward if each object in the initial state had been assigned the negative of whatever velocity it in fact had—an alternative velocity assignment that, as mentioned above, is just as probable as the assignment that was actually generated.

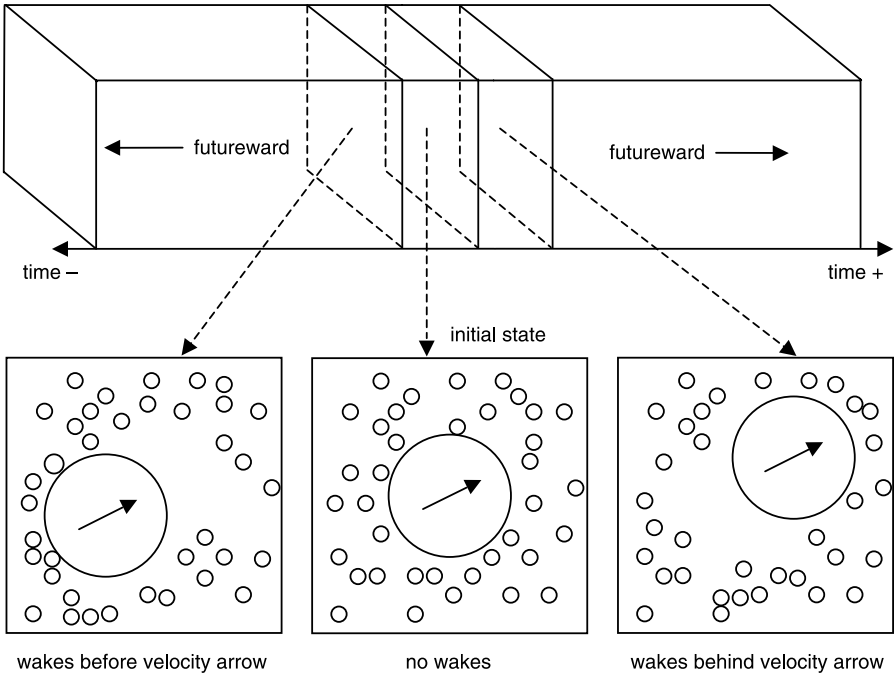


Figure 3.7
With time-symmetric physical laws and a time-symmetric initial state, the apparent futureward arrow bifurcates, pointing in both directions along the time axis from the initial state (schematic).

Thus, as we look along the time axis on either side of the initial state, the seemingly forward direction of time is simply the direction *away from* the initial state. Just as terrestrial *up* (the spatial direction from which things fall) once seemed to be a single direction—back when the earth was thought to be flat—but later turned out to be all directions pointing away from the center of the earth, so too does *futureward* (the temporal direction in which wakes trail behind, etc.) turn out (if we contemplate the application of time-symmetric physical laws before the initial state) to consist of both directions away from the initial state. (Boltzmann, who pioneered the field of statistical mechanics, proposed this analogy with regard to exponentially unlikely fluctuations of entropy that would randomly result in locally "time-reversed" regions [Boltzmann 1996, originally 1897].)

Thus, neither the physical laws nor the initial state itself needs to have any time symmetry to account for the seeming time asymmetry of (for ex-

ample) the temporal direction in which wakes trail behind—because that very phenomenon is time symmetric after all. The futureward direction (with respect to wakes, etc.) goes one way on one side of the initial state, and the other way on the other side—an entirely symmetric arrangement with respect to the two directions.

Still, there must be something about the (albeit time-symmetric) initial state that is special with regard to time. After all, suppose we looked at some other cross section of the artificial universe's spacetime and decided to designate *it* as the initial state. That is, we specify the objects' initial positions and velocities exhaustively: we make a list of the positions and velocities we find in the newly designated ''initial'' state, and specify those as the initial positions and velocities, rather than generating them at random. Then we generate the rest of spacetime, in both temporal directions, by applying physics to the new initial state. We would not, of course, then see the futureward direction radiating both ways from that cross section instead of from the original initial state. Rather, we would see a universe whose contents are identical to the previous one's contents. Merely moving the label ''initial'' to one of the other states does nothing to change what's already depicted in the spacetime of the universe.

As remarked earlier, the low entropy of the initial state does not, by itself, explain its privileged status. Other nearby states have entropy only slightly higher, with the giant disks just slightly slowed, and the mini-disks just slightly sped up. Furthermore, we could specify an alternative version of the disk-universe in which the initial state's random velocity assignments have slightly higher entropy than in the initial state of the first disk-universe (say, the giant disks' average speed exceeds the mini-disks' average speed by a lesser margin in the alternative universe than in the original). Yet the apparent arrow of time would still radiate in both directions from the alternative universe's initial state. Or conversely, we could specify yet another alternative version, in which the randomly generated initial state has *lower* entropy than in the first universe (with more of a gap between the giant disks' and mini-disks' initial speeds). We can identify a subsequent time-slice of that alternative universe whose entropy has increased so that it equals the entropy of the first universe's initial state. Yet the apparent arrow of time does not bifurcate at that subsequent time slice of the new universe. Thus, there is no specific entropy level that makes the apparent arrow of time radiate from a given state.

What distinguishes a randomly generated initial state from any futureward state is that in the initial state, the positions and velocities of the individual objects are uncoordinated, because the states were assigned *independently* (or rather, the states were generated almost independently, except for overlap avoidance). Looking at any given disk tells you nothing about the positions of any of the other disks (except that they don't overlap the given disk).

Once we start applying physics to the initial-state snapshot to generate new states (in either direction along the time axis), the physical laws prescribe interactions among the colliding disks. The disks thereby affect one another, and thereby come to bear information about one another. In particular, an ensemble of miniature disks, pushed out of the way by a giant disk, forms a wake that momentarily "remembers" the disk's recent passage. Looking at the wake alone, you can infer an elevated probability that a giant disk is nearby; looking at the giant disk alone (and its velocity vector), you can conclude that there is probably a wake in the direction the disk came from.

More subtly, even in a subsequent universe-state in which the wake at a given location has dissipated, the entire configuration of the giant and mini disks—their positions and velocities—continues to bear full information about that wake. It is because of that information, that coordination, that if we were to run the universe backward from that subsequent state—applying physical laws to generate each successive previous state rather than each next one—we would find mini-disks "conspiring" to get out of the way in advance of the passage of each giant disk. That conspiracy would be unlikely to the point of virtual impossibility if it were to unfold from applying physical laws to a universe-state generated at random, because very few possible configurations lead to that spontaneous evacuation compared to the exponentially vast number that do not.

Entropy is important to time asymmetry, because without the low entropy of the initial state, conspicuously asymmetric phenomena such as wakes would not occur at all. But it is not the initial state's entropy level per se that distinguishes it as the point from which the apparent arrows of time emanate; rather, that distinction comes from the lack of coordination among the constituent objects' states. The postulated independent (or almost-independent) generation of the objects' initial states is one way to ensure (with near certainty) that lack of coordination; alternatively, a sim-

ple, deterministic initial state—say, with particles uniformly spaced, with uniform initial speeds, and some periodic pattern of directions—would also lack such coordination, and would also serve as a state from which the apparent arrows of time emanate. (The disk-world thought experiment is a different way of rehearsing many of the arguments about entropy and reversibility that Albert 2000 presents more thoroughly and more precisely, yet delightfully accessibly.)

Thus, as in the mirror paradox in section 1.2.3, we find that the underlying laws can be symmetric after all and still produce the apparent asymmetry that we perceive—despite the fact that the phenomena we perceive can be reduced to the underlying laws and that symmetric underlying laws cannot give rise to asymmetry. Nor, however, is the perceived asymmetry merely illusory—rather, as in the mirror paradox, it is misinterpreted.

The time asymmetry in the events we perceive is not really between one time-direction and the other, the positive- and negative-designated directions along the time axis. Rather, the asymmetry is between the direction *away* from the initial (uncoordinated) state and the direction *toward* the initial state. And that asymmetry is itself symmetric with respect to the positive and negative temporal directions (since the *futureward* arrow bifurcates, pointing away from the initial state in both the positive and negative directions). Because of that symmetry with respect to the positive and negative temporal directions, the perceived events are not inconsistent after all with underlying laws that are symmetric with respect to the positive and negative directions.

To emphasize how the apparent arrow of time depends on the lack of coordination within the initial state, and depends on the developing correlations within subsequent states, consider the following thought experiment. We assign the initial state as above and run physics on it in both directions along the time-axis to generate the content of all spacetime. We then modify the resulting spacetime content as follows. We look at a snapshot of the whole universe at some point well futureward (in either direction) of the initial state (say, the snapshot in fig. 3.8b). We alter that snapshot by randomly assigning new directions to the giant disks (but leaving their speeds the same). We do not alter the mini-disks at all. Then we run physics in both time-directions from the altered snapshot to generate a new content for all spacetime.

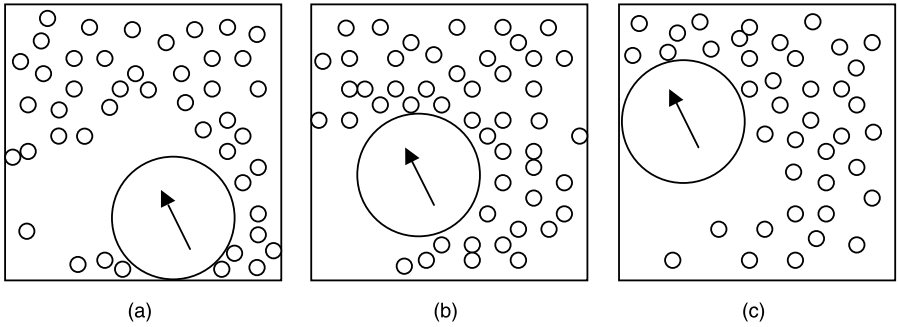


Figure 3.8
In (b), we randomly assign the giant disk the new direction depicted by the arrow. We then re-calculate previous state (a) and subsequent state (c) (schematic). Like (c), (a) now seems futureward from (b)—the wake in (a) corresponds to the disk's position in (b), rather than vice versa.

Looking at the new content in the futureward direction, we see a future that, although different from the original one, looks just as plausible as the original (except for a transient anomaly right at the time of the altered snapshot—due to the giant disks' reassigned velocity vectors, the wakes are momentarily uncoordinated with the giant disks' new directions). Wakes still trail behind the giant disks (except for the brief anomaly), just as in the original future (fig. 3.8c).

But if we follow the new content in the pastward direction from the altered snapshot (that is, back toward the initial state), we see something strange. Instead of seeing an obviously backward-running movie in which the wakes precede the giant disks (which is what we'd see if we watched backward from the unaltered snapshot), we see what looks like a plausible forward-running movie. The exquisite coordination required for the movie to appear to be running backward (as the mini-disks seem to magically scramble out of the way in anticipation of the giant disks' imminent passage) has been destroyed by the reassigned velocities of the giant disks. Without that coordination, the giant disks just plow through the mini-disks, pushing them aside in the usual fashion (fig. 3.8a).

Hence, the apparent arrow of time now points in both directions away from the altered-snapshot state, instead of pointing in both directions away from the initial-snapshot state. Yet we have not decreased the altered snapshot's entropy by randomizing the giant-disk velocities there. It is the

absence of coordinating information, rather than just low entropy, that distinguishes the state that the apparent arrow of time emanates from.

Of course, the alteration of the snapshot does not correspond to anything that could be accomplished from within the disk universe by the motion of the disks themselves. On the contrary, the laws of physics, together with a given initial state, determine the static, unchanging content of all spacetime, and of all its constituent snapshots. The point of the thought experiment is merely to underscore the relationship between the coordinating information among objects and the apparent arrow of time, by showing how that apparent arrow would be modified if (impossibly, given the physical laws) the coordinating information were to be compromised.

Let's return to the real universe. Although things are much more complex here, the basic lessons above regarding time symmetry still apply.

- All our well-confirmed physical laws are indeed time symmetric, such that the function that determines the immediate future on the basis of the present is a mirror image of the function that specifies the immediate past on the basis of the present.[8] If you look at a movie of an elementary particle's behavior in isolation, or a simple interaction between two elementary particles, you cannot tell if the movie is being shown forward or backward—either way, the same physical law is followed.
- Nonetheless, when we look at large ensembles of particles, their behavior shows a conspicuous arrow of time. There are myriad examples, all analogous to the wakes in the artificial universe. A physical wake in water—left behind by a boat, for instance—is one such example. A waterfall is another—or an egg rolling off a table and shattering, or air rushing out of a pressurized tank through an open valve, or the formation of footprints in the sand, or the imprinting of a scene on the film in a camera, or the behavior of people or other organisms. Indeed, any conspicuous macroscopic effect amounts to a wakelike record of its cause. A video of any such example, if shown backward, would be quickly, obviously distinguishable from one shown forward. Still, if you were to look at the behavior of each elementary particle and its interaction with its immediate neighbors, you

8. Again, as mentioned above, the real world's requisite mirror-imaging turns out to involve more than just reversing velocities. A couple of other elementary properties have to be swapped as well.

would see the same physical laws followed in the backward-running movie as in the one running forward.

So, for example, even the reassembly of a shattered egg, if we watch a movie of it in reverse, follows ordinary physical laws: by a seemingly exquisite coincidence, the surrounding molecules just happen to jiggle the eggshell fragments in such a way as to propel them together with just the right orientation and energy to reconstruct the bonds that hold the shell together, and so forth. Of course, it is not really an exquisite coincidence, but rather an exquisite coordination, made possible by the (forward-running) interactions by which those surrounding molecules—slightly heated by the egg's impact—''remembered'' or ''recorded'' the interaction via the details of the increased molecular jiggling that constitutes the increased heat. This detailed recording lasts forever—at any point in the distant future, applying physics backward in time would return us to the event of the egg's unsplattering.

But this recording is unobvious, distributed in the exact details of the states of all the affected particles. The recording would become apparent only if—impossibly, except in a thought experiment—we could actually watch the ensemble running backward in time. In contrast, the stain left by the splattered egg—which we can readily observe, but with no clue about the detailed state of its constituent particles—constitutes a more obvious recording of the event of the egg's fall, one whose significance as such a recording can be appreciated without having to play the events backward in time. Like the artificial-universe wakes, this recording is transient (though less so than the wakes). Eventually (even if it takes many years), all obvious remnants of the egg dissipate, unlike the fully detailed record, which is permanently—but inaccessibly—conserved.

A wake, or a footprint in the sand, or a photograph, is a similarly obvious memory or recording of the event that formed it; again, we do not need to run things backward to be able to infer what event was likely to have produced the recording. Indeed, any artifact that is designed to remember or record—or any organism that evolves to do so—makes use of some real-world physical process similar to the process that forms wakes in the artificial universe. Some processes create records that are less transient than others, or more readily manipulable, or more compact, or more rapidly accessible. The machinery for storing and retrieving information in a digital

computer's memory, or in a human brain, is vastly more sophisticated than a wake, but it still harnesses the same sort of process as a wake: a process whereby objects of mutually uncorrelated macroscopic states interact in such a way as to create macroscopic correlations, in the "futureward" temporal direction—that is, the direction away from a distinguished uncoordinated state. Thus, in particular, the (metaphorical) Cartesian Camcorder that underpins our consciousness (sec. 2.2 above) records the past but not the future, without violating the time symmetry of the underlying physical laws.

What, then, is the uncoordinated, low-entropy state from which the apparent futureward arrow points? For our universe, an obvious candidate is the moment of the big bang, which is distinguished in several striking ways: the content of the universe was maximally compressed, and maximally uniform, in space.[9] The physicist Stephen Wolfram (2002) speculates that the initial state of the universe might be as simple, as orderly, and as compactly describable as the laws of physics themselves. In that case, initial positions and the like may well be assigned at random, uniformly over space—or even deterministically, in a simple, regular, uniform, spatially symmetric pattern.[10]

Uniformity implies high entropy in a gas, or in a Newtonian artificial universe like the ones discussed above. In those domains, interactions from a random starting configuration tend toward increasingly uniform dispersal. But given a sufficiently enormous density of matter and energy,

9. Some physicists (Andrei Linde et al., as discussed, e.g., in Greene 2004) have developed a model according to which the big bang does not occur at or near the "initial" state, but rather is one of many big-bang events in a much wider universe. Still, if there were some distinguished (uncoordinated, low-entropy) state, even if long before our own local big bang, then that distinguished state would account for an apparent arrow of time, despite time-symmetric physical laws.

10. Under Newtonian mechanics, there are forms of spatial symmetry that, once established, can never be violated, given the spatially symmetric laws of motion. But under quantum mechanics, discussed in the next chapter, a spatially symmetric placement of particles, with identical velocities, leads to a probabilistic distribution of states (literally or figuratively probabilistic, depending on your interpretation of quantum mechanics, as the next chapter addresses). Individual members of this distribution can be spatially asymmetric (although, symmetrically, there are other members with opposite asymmetries).

the gravitational tendency toward clustering (and the concomitant warping of spacetime itself, according to general relativity) makes early cosmic uniformity a very rare, precariously balanced configuration. Hence it is a low-entropy state in the sense that almost all randomly generated configurations quickly lead to highly clustered, *non*uniform states. Even slight deviations from uniformity lead to the collapse of matter into galaxies, stars, planets, and black holes.

This ever-increasing gravitational concentration constitutes a progressive increase in entropy, a progressive winding down into an eventually boring equilibrium, as with the gradual slowing of the giant disks in the artificial universe discussed above. (Luckily, though, the real universe takes considerably longer to wind down, and is much more complex and interesting in the interim.) And just as those high-velocity disks, when they interact with mini-disks and slow down, create temporary local decreases in another aspect of entropy (namely, the formation of lower-entropy wakes momentarily disrupts the higher-entropy spatial uniformity of the mini-disks), so too does the concentration of matter in the real world cause local entropy decreases—most impressively, by forming solar systems whose radiant energy and larger atoms sometimes lead to the organization of matter into life.[11]

Still, it is the lack of coordination among the particles in the universe's initial state—rather than that state's entropy level per se—that would explain the apparent futureward arrow pointing only away from the initial state, as evidenced by wakes, waterfalls, footprints, photographs, and other such phenomena. And if the laws of physics are time symmetric, then it is a straightforward deduction that the futureward arrow points in opposite temporal directions on opposite sides of the initial, uncoordinated state.

But does this mirrored half of the universe on the other side of the initial state really exist? Of course we can have no evidence either way. A universe temporally truncated at the initial state would look just the same to us, in our own time-slice, as would an untruncated universe. For that matter, though, a universe truncated ten minutes ago—a universe that just sprang into being with the then-present time-slice, with no past before it, although spurious memories of the past are embedded in that time-slice—

11. See Greene 2004 for an admirably accessible discussion.

would look the same to us now as an untruncated one. The most we can say is that if we take the physical laws that appear to hold without exception, we can apply them pastward and extrapolate further back than ten minutes ago—or not. And we can keep going, extrapolating further back than the "initial" state, as well—or not. Admittedly, truncating at the big bang is less arbitrary than truncating at ten minutes ago, but neither truncation is dictated by the laws themselves.

Our current knowledge of the universe's physical laws does not let us get closer than a small fraction of a second to the big bang. But if there is indeed a distinguished initial state, then presumably we will eventually know how to extrapolate all the way back to it. And if the ultimate physical laws are indeed time symmetric, as the well-established approximations are so far, then they must indeed apply on both sides of any such state.

I briefly return to related metaphysical questions in section 8.1. For now, though, suffice it to say that the reality of the other temporal half of the universe doesn't much matter to us, except insofar as it resolves the paradox of a seeming temporal asymmetry arising from temporally symmetric physical laws. Apart from that resolution, the other universe-half has no practical importance. In particular, the very physical principles from which the other temporal half can be inferred in the first place also ensure that communication or interaction between inhabitants of the two halves is flatly impossible, because the establishment of macroscopic correlations (which is what such communication or interaction would consist of) proceeds only futureward (in both directions) from the initial state itself. The other half is real to us only insofar as we are able to infer its existence from whatever the ultimate physical principles turn out to be.

Let us return now to the postponed question of the apparent flow of time—the seeming existence of only the present moment, but an ever-changing present, moving forward in time. The foregoing considerations suggest that insofar as the time-symmetric laws of physics implement events that exhibit an apparent arrow of time, it is a bifurcated arrow, pointing in both directions away from the initial, uncoordinated state. But the beginning of this chapter argued that the laws of physics do not in fact implement any *flow* of time, any flow of the present moment in the direction of the apparent arrow. Rather, the content of the universe is just sitting statically in spacetime.

There is, to be sure, an ordering relation that proceeds in both temporal directions from the initial state. It is the order in which a maximally simple description of the universe can be used to derive, to compute, the entirety of spacetime. The description consists perhaps of a few straightforward (time-symmetric and space-symmetric) laws, plus a (crucially to this discussion) uniform, simply described distinguished (initial) state.

But there is no reason to suppose that there is any process external to this universe that actually implements the computation by which the whole content of spacetime can be derived. (Sec. 8.1 below briefly addresses the question of where, then, the universe might ''come from.'') Such a process would involve an additional dimension, a metatime, with respect to which some pointer to the current time is in motion, pointing to a different ''present'' at different metatimes (even though our physical laws involve no such metatime).

More to the point, even if there were such an external process, it could not be noticeable from within the universe and hence is superfluous in explaining why there seems to us to be such a flow. Not only would the supposed flow seem the same to us in the absence of any such external process, it would seem the same to us even if an external process were instead running backward—that is, opposite the futureward direction, starting with an (exhaustive, not at all compact) description of some distant-future state. The content of the universe's spacetime would still be the same regardless of the order in which the time-slices are (externally) viewed or generated. Viewing the backward generation step by step, an external observer would not see you remark ''!drawkcab gninnur si emiT !tihs yloH''. Rather, if you happen to address the subject at all, you would merely be seen to say (and think) ''.drawrof gninnur si emiT'', as usual, and you would say it for the same reasons as usual.

What, then, creates this impression of a flow of time? As mentioned earlier, to move a minute into the future, you need only wait a minute, and there you are. Not so with moving into the past. But why the difference? Could you not, after all, just wait -1 minute and then find yourself a minute earlier? If we define waiting n minutes simply as comparing your present state to your state n minutes from the present, then it is indeed just as easy to wait -1 minute as it is to wait $+1$ minute: you simply compare your present state to your state in -1 minute, that is, a minute ago.

Surely, though, comparing your present state to your state a minute ago does not correspond to what we think of as waiting from now until that other state. The notion of waiting *n* minutes thus smuggles in some additional meaning beyond a mere comparison between the present state and the state *n* minutes from the present. The additional meaning is intuitively clear: to wait entails in part that at the end of the wait, you can remember the beginning of the wait and the events in between.

Thus, the version of you at 10:00 can anticipate, but cannot remember, what it will be like to be at 10:01 the same morning. The version of you at 10:01 can remember what it was like at 10:00, including the anticipation at 10:00 of being at 10:01, but cannot (correctly) anticipate being at 10:00—because there can be no version of you at 10:00 that remembers an anticipation at 10:01 of being at 10:00. This subjective ordering of time, based on the sequence of inclusions of memory, points in the same temporal direction as the apparent futureward direction that is exhibited by diverse physical phenomena such as wakes, waterfalls, and splattering eggs—and not coincidentally. As discussed above, any mechanism for memory harnesses some such physical process to make its recordings, so it too points in the same direction.

From the point of view of physics, with its static spacetime, there is merely a collection of different versions of you, thinking and saying different things at different moments. Nothing ever designates one of those moments as the present and then changes the designation, sliding it futureward along the time axis, implementing a flow of time. But there is a sequence defined by the inclusion of memories. The version of you at one moment has memories of the versions of you of many previous moments—including memories of some of those versions remembering *their* previous moments' versions. And no version remembers any of the future moments' versions (for the reasons discussed above).

This sequence of inclusion of memories does define a sort of metatime against which differences in actual time can be measured—differences indicated by clocks ticking and other physical processes. This metatime is not, of course, a separate physical dimension, nor does it reflect any external process of sequentially generating or scanning the time-slices of the universe. As a property of our cognitive machinery, it is fundamental to our brains' perceptions, but not to our universe's physics.

Still, the physical and the psychological are easy to confuse here (just as, in sec. 1.2.3's mirror paradox, the psychological preference that interprets a reflection as left–right reversed but not top–down reversed was easy to confuse with a seeming physical asymmetry in the mirror's treatment of the horizontal and vertical axes). We remember the past, but anticipate the future—not because of any time asymmetry of physical laws, but rather because of information that interacting objects exchange, in the temporal direction that points away from the distinguished state in which objects are mutually uninformative. The distinction between memory and anticipation imposes a subjective ordering on the moments we experience, creating the illusion that time itself flows in sequence according to that ordering. In reality, though, the laws of physics prescribe instead a collection of different temporal versions of ourselves, some remembering others, but all sitting statically in spacetime, with no flow of time at all.

For most practical purposes, of course, we have no reason to stop speaking and thinking in terms of a flow of time, an ever-advancing present moment—just as, for most everyday purposes, we continue to think and speak of the sun rising and setting, even though its apparent trek across the sky is really due to the rotation of the earth. It would be pedantic in these cases to reject the benign convenience of treating superficial appearances as real. But when our goal is to understand, not just to muddle through the day, we need to draw distinctions more carefully.

3.3 Summary

The contents of spacetime are static. Time does not flow, any more than space does. There is no metatime dimension with respect to which time could flow. There is no pointer sliding along the time axis of spacetime, designating an ever-changing present moment, any more than there is such a designation for each spatial axis. An imaginary observer external to the universe might view successive time-slices in the forward direction, or backward, or view the entire collection simultaneously and statically. From any such view, the content of spacetime is the same—including the part of the content in which you and I perceive and remark that time seems to flow forward.

The content is dictated by a distinguished state and a set of physical laws of motion or state-change. Even if there were some flow of time (forward or

backward), it would make no difference to the events in the universe. In particular, it would make no difference to our perception of the apparent forward flow of time, for any such perception is itself among the universe's events, determined by the distinguished state and the physical laws. Whatever we are actually responding to when we think we perceive that flow is, therefore, something else.

Moreover, the physical laws of motion and state-change are time symmetric: they treat the past (as a function of the present) as a mirror image of how they treat the future (as a function of the present). A movie of an individual particle, or of an individual interaction between two particles, appears to follow the same physical law whether the movie is running forward or backward. Yet if we watch a large ensemble of such interactions, what we see if we watch the movie backward—for example, a wake that spontaneously forms in anticipation of an object's imminent passage—is unlike anything we would ever see when we watch it forward. The contrast is paradoxical, because time-symmetric laws that implement time-symmetric interactions could not thereby produce an ensemble of interactions that is time asymmetric, treating one temporal direction differently than the other.

But the production of wakelike ensembles is exhibited even by simple artificial universes with time-symmetric laws. The paradox of an apparent asymmetry arising from symmetric laws is resolved by the observation that the seemingly futureward direction (with regard to the behavior of wakes and the like) is the direction (forward or backward) away from an "initial" state—a state distinguished not just by low entropy, but by the lack of coordination among distinct objects' states. That futureward direction bifurcates at the initial time-slice, and is therefore symmetric after all with respect to the two temporal directions.

In our universe, too, the uniformity of the seemingly futureward direction can be explained by supposing that there is a distinguished state (plausibly, at the moment of the big bang) in which individual objects' states lack the sort of coordination that could establish reverse wakes. On that supposition, the initial state itself may be compactly describable (say, in terms of a simple random distribution or in terms of a uniform spatial arrangement), as are the laws of state-change. There is an Occam's-razor principle that prefers to postulate succinct, uniform laws if they accord with our empirical observations—rather than, say, laws that exhaustively

propose quintillions of arbitrary, individually specified miraculous exceptions here and there. That same Occam's-razor principle similarly prefers a distinguished-state description that is compact and uniform (if it accords with our observations)—rather than one that must exhaustively specify the state of every particle, as would be necessary to rig a miraculous ''conspiracy'' that could produce, for instance, a spontaneous, anticipatory wake when watching events backward if we could not instead specify a simple initial state whose futureward interactions create the conspiracy.

Starting with the distinguished, uncoordinated state, and proceeding in both temporal directions, interactions between particles (according to physical laws) give distinct particles information about one another. Some resulting ensembles, such as wakes, constitute obvious macroscopic recordings of the events that caused them. Explicit human memory is one of myriad phenomena that harness wakelike processes to make and view recordings. As such, human memory partakes of the futureward directionality of the underlying processes.

Among the many versions of you collected in spacetime, each of course has its own ''now,'' just as each has its own ''here''—that is, each version has a particular location in time and in space. The puzzle is that ''now,'' unlike ''here,'' seems to flow ever forward, whereas physical laws describe no such flow. But the illusion of this flow is explicable: futureward versions of you remember pastward versions and events (including remembering that the pastward ones remember further-pastward ones), but not vice versa. This temporal directionality of memory and anticipation defines a flowlike ordering of moments, even though there is no actual flow. The blurred distinction between your thoughts and your (slightly subsequent) consciousness of your thoughts (chap. 2) may help create the illusion, at each moment, that the current version of you is the same as the (consciously remembered) previous version, and that it has just been nudged forward in time.

Conceiving of the universe as having static spacetime may be acceptable from a physicist's point of view, especially if the above considerations succeed in reconciling it with the appearance that time flows (and that it appears to flow exclusively in one direction). Still, from our day-to-day, commonsense point of view, static spacetime wreaks havoc with any notion of making meaningful choices for the sake of the future. If it's all just

sitting there already, the future as much as the past, then what room is there for our choices to *change* anything?

As mentioned in chapter 1, this consideration is one of the tendrils that link together far-flung ideas, entangling our basic conception of the physical universe with our notions of how we ourselves operate from moment to moment. Just as a round-earth proponent must explain why those on the other side of the planet do not find themselves falling away from the ground (or else must concede, mistakenly, that they do), a static-spacetime proponent must explain how genuine choice is possible in such a universe (or else must mistakenly abandon the notion of such choice).

Chapter 5 addresses in depth the question of choice in a mechanical, deterministic, static-spacetime universe. First, though, the next chapter examines perhaps the most bizarre challenge yet to the notion of such a universe—the phenomena of quantum mechanics. Although these phenomena indeed reveal a universe that is wildly unlike what either our common sense or our pre-1900s science would suggest, I argue that its peculiarity turns out to be radically different from the interpretation that physicists initially proposed, and that later trickled down to much of the lay public.

4 Quantum Certainty

4.1 The Quantum Paradox

The preemption of classical physics by quantum mechanics is widely regarded—not least by many physicists themselves—as a fundamental retreat from the ideal of a mechanical, clockwork universe. Physics, which was once the best exemplar of the mechanical paradigm, now seems to be its most formidable detractor.

The well-known apparent nondeterminism of quantum mechanics is the least of its oddities; probabilistic laws still afford a straightforwardly mechanical model. Far stranger is the apparent observer-dependency of nature:

- Of several states that a particle might be in, it turns out that all may coexist. It is as though there are several simultaneous versions of the particle, each in a different state, as is shown (statistically, over many trials) by the different versions' mutual interference.
- Bizarrely, however, whenever we observe the particle in this so-called *superposition* of states, we see just one version in just one of the previously coexisting states. And thereafter, the previous superposition vanishes, with no further trace of the other superposed states, as though they had never been present in the first place. (*Which* of the states we observe is unpredictable in principle—hence, the apparent nondeterminism.)

If even inanimate physical objects are not mechanical in nature, and especially if their reality depends somehow on us observers, then surely we observers are not plausibly just mechanical. Not surprisingly, then, the defection of quantum mechanics from the clockwork camp has been seized on as a vindication of the view that consciousness is not merely a particular

example of mechanical computation, but rather is something that is special all the way down to its underlying substance.

Several decades ago, however, the physicist Hugh Everett (1957) noted that the mathematics of quantum mechanics does describe a straightforwardly deterministic system after all. There are quantum superpositions, but we observers are among the physical objects that can be in a superposition of states. Each such superposition pairs a particular state of the observed object with the corresponding observer-state, so each superposed version of the observer observes only one of the superposed states of the object, even though the superposition persists. Because there are multiple superposed versions of the observer too—each observing just one superposed state—the fact that (each version of) an observer sees only one such state does not imply that the superposition has (randomly) collapsed to just one such state.

The very concept of a conscious observer in a superposition of states—some versions observing one thing, other versions observing another—is anathema to the view of consciousness as a unitary, transcendent, extraphysical, extramechanical phenomenon. But if consciousness is just an ordinary physical process, its superposition is no stranger than that of any other physical process (which, admittedly, is strange enough; quantum superposition, however, is a fundamental property of physics, its reality established empirically beyond any trace of a doubt).

Thus, the common argument from quantum-mechanical randomness to extraphysical consciousness gets things exactly backward. When a particle that had been in a quantum superposition is observed, the seeming collapse of the superposition into a single random state stems from failing to conceive of the observer herself as a physical object subject to quantum superposition. This failure smuggles the notion of nonmechanical consciousness into the very interpretation of quantum mechanics. If the smuggling goes unnoticed, it creates the false impression that quantum mechanics itself provides independent evidence for nonmechanical consciousness.

Here again, disparate ideas entangle, as when recognizing the roundness of the earth compels us to revise the physics of up and down, lest it seem that our far-side counterparts must stand on their heads. Commitment to a nonmechanical concept of mind distorts our interpretation of physics; thus distorted, physics seems to support a nonmechanical concept of mind. To

stand ourselves right-side up, we must (at least tentatively) entertain the mechanical paradigm for both mind and physics, and then evaluate the coherence of the proposal. Chapter 2 made the case for viewing consciousness as mechanical (provided that other problems can be resolved, in this chapter and the following ones). This chapter makes the case for deterministic, clockworklike quantum mechanics (provided that consciousness can be seen as mechanical).

But by any interpretation, quantum mechanics is far too strange to be understood just by a narrative description like the foregoing. Rather, we need a formalism to precisely render the underlying ideas. This chapter first highlights the seeming paradox of quantum mechanics, then presents a simple formal model that, using little more than high-school mathematics, illustrates Everett's solution to the quantum paradox—a solution that rescues the mechanical paradigm, restoring determinism and observer-independent reality to quantum physics, and restoring physical ordinariness to conscious observers.

4.1.1 The Double-Slit Experiment

The classic double-slit experiment highlights the quantum paradox. We aim an electron at a pair of adjacent, narrow slits in a barrier (imagine this happening in just two dimensions). Beyond the barrier lies a backdrop with a row of adjacent electron detectors. Each detector has the same width as each of the two slits; the distance between the two slits is much greater than this width. If the electron passes through the barrier via the slits, we find that one (and only one) detector soon registers the electron's arrival at the backdrop.

Suppose we block one of the two slits and conduct many trials of this experiment, graphing the distribution of electron-arrivals at the various detectors. Not surprisingly, the graph shows a smooth curve with a peak opposite the unblocked slit; this curve shows that the electron tends to continue straight ahead, perhaps diverting slightly to one side or the other, but larger diversions are less probable.

If instead we unblock the other slit, then, of course, the distribution curve has a peak opposite that other slit. If we conduct a number of trials, half with one slit blocked and half with the other blocked, the distribution curve is just the sum of the two single-slit curves. All of this is consistent with the electron's being a particle that is smaller than the width of each

slit and that passes through the currently unblocked slit if it happens to reach that slit.

But now, suppose we try many trials of the experiment with both slits unblocked *simultaneously*. Bizarrely, the distribution curve is not the expected sum of the single-slit curves. Rather, the curve shows an interference pattern. At some points along the backdrop, the frequency of an electron's arrival is not only less than what the sum of the single-slit curves predicts—it is even less than what either single-slit curve alone would predict!

The distribution actually seen over a large-enough number of trials must approximate the sum of the probability distributions of the individual trials. Hence, the distribution curve shows that by unblocking both slits together—by providing an additional path by which an electron might arrive at a certain point along the backdrop—we have somehow reduced (rather than increased) the probability of its arriving there on a given trial, compared to what happens if either one of the two slits is blocked.

This result is inexplicable if the electron indeed passes through just one slit or the other. If a given electron encounters only slit *A*, opening slit *B* could not reduce the likelihood of the electron's reaching a given destination through slit *A*. But the interference is just what we would expect if the electron were not a spatially localized particle, but rather an expansive wave—a wave that passes through both slits, creating typical wavelike interference on the other side of the barrier.[1] Indeed, the observed interference pattern accords quantitatively with the predictions of wave mechanics. The wave's amplitude at a given point corresponds to the probability (the probability is actually the square of the amplitude) that the electron arrives there, as seen by a detector at that point.

But this wavelike phenomenon raises an apparent paradox. If the electron spreads out in a wavelike fashion, why does the backdrop detect only a local, discrete arrival for each electron? Why does only a single detector react, rather than many adjacent ones? As noted above, the statistical distribution over a large number of trials warrants an inference about what occurs on *each* trial. We can thus infer from the statistical evidence that the electron passes through both slits on each trial. The universe thus

1. Wave mechanics—or just common sense—tells us that when two waves mesh together, there are points at which the crest of one corresponds to the trough of the other, so they can cancel out at that point, producing less of an effect there (or even none at all) than if either wave alone had been present.

seems to be playing hide-and-seek. Whenever we detect the electron, we see a localized particle. But when we do not observe it, the electron is a wave, passing simultaneously through two widely separated slits (widely separated compared to the size of the particle itself), and exhibiting interference on the other side.

We might seek to clarify the situation by shining a light source on the barrier to see the electron as it passes through. In that case, we unambiguously see the electron emerge from just one slit or the other. But then, the distribution curve, over many such trials, no longer shows interference. Instead, it simply equals the sum of the single-slit curves, just as we would expect if the electron were a particle rather than a wave.

4.1.2 The Interference–Observation Duality

Thus we have the fundamental, paradoxical duality:

- There are coexisting, mutually interfering states, so long as the states are not distinguished by observation. (In the double-slit experiment, there is a continuum of such states, propagating in a wavelike fashion.)
- Whenever an observation is made, only one of the superposed states is ever seen. (In the double-slit experiment, a conventional particle, much smaller than the wave, is all we see when we look.)

This is known as the quantum-mechanical *wave–particle* duality. A standard understatement of this duality is that an electron (or other physical entity) acts sometimes like a wave, sometimes like a particle. More strikingly, we have here an *interference–observation* duality: there are many superposed, mutually interfering states whenever we're not looking, but only one such state whenever we do look. (Heisenberg's uncertainty principle says, moreover, that no matter how precise an observation we perform, some superposition must remain. Indeed, the more precisely we measure a given attribute, the more superposition there is with respect to some other attribute.)

To see just how dramatic the interference–observation duality really is, consider John Wheeler's *delayed-choice* modification of the double-slit experiment (1983): we do not decide, until just after the electron passes the barrier, whether to collect the electron against the backdrop, or whether to pull the backdrop out of the way and observe which slit the electron came through (by using a pair of lenses, each focused on one slit):

- If we choose to remove the backdrop and make the observation, we find that the electron passed through just one of the slits.
- If we choose not to observe, the distribution we find (over many such trials) is once again consistent with the "particle" having passed, wave-like, through *both* slits on *each* trial, the two parts of the wave mutually interfering.

What, then, does the electron do when it reaches the barrier, prior to our decision whether to observe its path: does it then pass through one slit or both? It seems that the answer is determined *in retrospect* when the distinguishing observation is made, or when the electron instead reaches the backdrop.

4.1.3 Interpretations: Copenhagen and Everett

The standard interpretation of such phenomena in the early 1900s—the Copenhagen interpretation—shows the profound effect of this paradox on physicists' sense of reality. According to the Copenhagen interpretation, no physical phenomenon is real until it has been observed. Nothing real passes through both slits of the apparatus. Instead, there is a *potential* for a real particle to pass through either slit, but that potential is not realized unless the passing-through is observed; at that point, the particle settles, at random, into one of its potential positions. The potential itself is wave-like, exhibiting interference effects.

This interpretation does, indeed, accord with the fact that the particle cannot simply pass through just one of the slits (else the interference would not be seen statistically), and with the fact that passing through just one slit is precisely what the particle has done whenever we look. But the Copenhagen interpretation gains this accord at the price of denying the observer-independent existence of the building blocks of reality.

Thus, quantum mechanics seems to challenge not only the world's determinism, but the very objectivity of the world's existence. Indeed, the Copenhagen interpretation provides no way to express the state of universe as a whole, since a system's state is real only with respect to an external observer, and the universe as a whole has no external observer.

The Copenhagen interpretation exhibits the usual rigor of physics to say what happens to the world *between observations*. What occurs between observations is given in part by Schrödinger's equation, which governs the

(fully deterministic) propagation of a (wavelike) quantum state of the universe. This state is a superposition of many individual, sometimes mutually interfering states (such as the state of an electron being at one slit or the other). When an observation occurs, Copenhagenists insist that the superposition of states *collapses*, leaving just one member of the previous superposition. But Schrödinger's equation itself does not describe any such event as this collapse.

What's worse, the Copenhagen interpretation has no formal criterion for what constitutes an observation (hence no criterion for when the putative collapse occurs). Is the detection of a quantum event by a laboratory instrument an observation? John von Neumann—the physicist, mathematician, game theorist, and inventor of digital computers—showed (von Neumann 1955) that the same prediction is made whether one stipulates a collapse at that point, or whether, on the contrary, one regards the superposition as persisting[2] (so that the macroscopic instrument is itself in a superposition of more than one detection state). Von Neumann's conclusion: only when a conscious being observes the state of the instrument—and sees that it is unambiguously in one state or the other—does it become clear that only one outcome occurred.

Thus was von Neumann (of all people) led to conclude that human consciousness (of all things) plays a fundamental role in physics: conscious observation precipitates the collapse of the quantum superposition. Most physicists, unlike von Neumann, accept that inanimate observation suffices to bring about the collapse. Still, a number of eminent theoretical physicists have shared von Neuman's version of the Copenhagen interpretation—quantum mechanics' most profound departure from the mechanical paradigm.

However, there is an alternative interpretation of quantum mechanics that restores a mechanical understanding of the universe. Quantum phenomena such as the double-slit experiment show that, prior to observation, the superposed states have symmetric status. That is, no one of the superposed states is already the unique real one. (*Hidden-variable* theories try to deny this, but such theories are provably wrong; see sec. 4.3.4 below.) Logically, then, there are two ways to achieve this symmetry: either none of

2. More accurately, the same prediction is made only when some trace of the observation persists. See section 4.4 below for elaboration.

the superposed states are real, or all of them are. The Copenhagen interpretation says none of the yet-unobserved states are yet real. Everett's *relative-state* interpretation (Everett 1957) says all of them are real.

In Everett's formulation, the quantum collapse never occurs. Superposed states remain in superposition even after observation (whether by inanimate objects or by conscious observers). It remains to account for the *apparent* collapse—the fact that we see only one outcome of the quantum observation. Everett's crucial insight is that the deterministic Schrödinger formalism already predicts an apparent collapse, even as it denies an actual one.

According to the formalism, observing a superposed state results in different versions of the observer in different versions of the universe, each version of the observer seeing a different outcome to the exclusion of all other outcomes. (Of course, it makes no difference whether the observer is animate—as opposed to, say, a recording device.) Thus, versions of the observers themselves are in superposition. But they are mutually isolated, so each sees a seemingly unique outcome. Following Everett, I argue here that this interpretation (which may sound desperately implausible on its face) is in fact the far more parsimonious one—but it takes a formal model to demonstrate that claim.

In this chapter, I try to make sense of the quantum-mechanical universe. Often, the best way to understand something is to build an example of it. Hence, this chapter builds a universe, a qualitative model of quantum mechanics (just as the previous chapter built some simple Newtonian universes to explore time symmetry). That is, I define a universe whose physics are quite different from (and much simpler than) our own world's. And I demonstrate that this universe exhibits an interference–observation duality analogous to that of real physics. (We can call this model *quantish physics*.) The analogy runs deep enough to support a comparison of the Everett and Copenhagen interpretations with respect to the quantish model; this comparison can help elucidate the interpretation of real physics.

I present three artificial universes: U1, U2, and the quantish-physics model U3. The first of these universes, U1, has straightforwardly classical mechanics. U2 attempts to incorporate quantumlike uncertainty in its physics, but fails in instructive ways. Finally, the quantish-physics model, building from the U2 attempt, succeeds in reconstructing the fundamental quantum interference–observation duality.

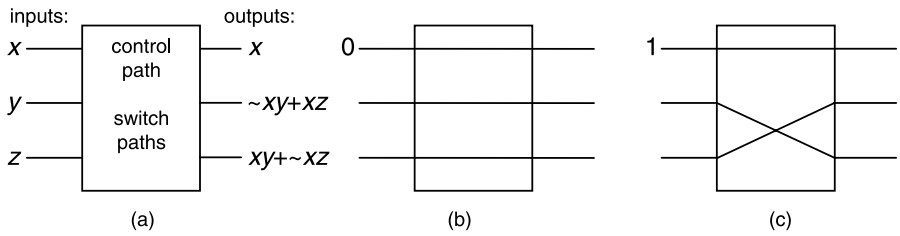


Figure 4.1
A Fredkin gate. In (a), Boolean-algebraic notation denotes the wires' respective outputs. For instance, $\sim xy + xz$ means *not-x and y, or x and z*. Thus, it is as though a 0 on the upper wire sets up the circuit in (b), whereas a 1 sets up the circuit in (c).

4.2 Illustrating Quantum Mechanics with Artificial Universes

4.2.1 U1: Configuration Space for a "Classical" Universe

Let us define a universe consisting of a circuit built from Fredkin gates (Fredkin 1982). A Fredkin gate has three binary (0 or 1) inputs and outputs. Each output computes a boolean function of the inputs, as specified by figure 4.1a. But the gate is more easily understood as having a *control path* going across the top of the gate, and two *switch paths* below. If the first input (the control input) has a 0, then the second and third inputs (the switch inputs) simply propagate to the second and third outputs (respectively), as suggested by figure 4.1b. If instead the control wire has a 1, then the two switch wires "cross," so the second input comes out at the third output, and vice versa (fig. 4.1c).

The control wire simply propagates its input to its output. All three paths through a gate impose a delay of one time unit between the appearance of an input value and its propagation to the corresponding output.[3] Fredkin gates, unlike some logic gates, do not allow fan-in or fan-out. Rather, each output must connect to exactly one input.

Fredkin gates are *computationally universal*. (Loosely speaking, their universality means that any logic circuit that can be built at all can be built using only Fredkin gates. Other gates, such as NAND gates, are universal as well.) Fredkin gates have the further property of *conserving 1s and 0s*—that is, the number of 1s (or 0s) that leave a gate equals the number that entered

3. In Fredkin and Toffoli 1982, delays occur in the wires rather than in the gates, but that difference is unimportant.

the gate one time unit earlier. Hence, the total number of 1s (or 0s) coursing through the circuit remains constant.

For a given universe (that is, a given Fredkin-gate circuit), one might represent the state of the universe at a given time by listing, for each wire, whether that wire has a one or a zero. Hence, the state can be represented by a vector $v_1, \ldots, v_n$ where v_i is 0 or 1 according to the state of the ith wire, and n is the number of wires in the universe. (A wire is defined to go from an output to an input. A gate's output wires are distinct from its input wires.)

Alternatively, because Fredkin gates conserve ones and zeros, we can index the world-state the other way around: for each 1—think of 1s as particles—we can say at which wire it currently resides. (To specify which wire a particle is at is to fully specify the particle's position; no gradations of position along a wire are recognized in this model.) We can depict such a particle-indexed state geometrically. If the universe has k particles, we can define a k-dimensional space. Each dimension has discrete coordinates ranging from 1 to n (the number of wires in the universe). For a given point $(p_1, \ldots, p_k)$ in this space, the point's ith dimension says which wire the ith particle is at. Call this space *configuration space*.[4]

A single point in configuration space thus represents the entire state of the universe, specifying the position of every particle. Rephrasing the physics of this universe in terms of configuration space, we get a rule for moving from one point in this space to another at each unit-time interval. Figure 4.2 illustrates this formulation. Suppose gate g appears in the Fredkin circuit defining our model universe (the rest of the circuit is not shown). Suppose for now that there exist just two particles, p_1 and p_2. Particle p_1 appears at g's control wire, p_2 at g's upper switch wire. Figure 4.2b shows the configuration-space point s_0 that designates this state of the universe. At the next time unit, the state of the universe becomes s_1. In that state, p_1 has moved to w_{1a} and p_2 has crossed over to w_{3a}.

The configuration-space representation is equivalent to, but more cumbersome than, the more obvious wire-vector representation. But in the fol-

4. Configuration space is analogous to *phase space* in real-world classical physics. For a system with k objects, phase space has $6k$ dimensions: three dimensions for each particle's position and momentum. Thus, a single point in phase space specifies the position and momentum of every object.

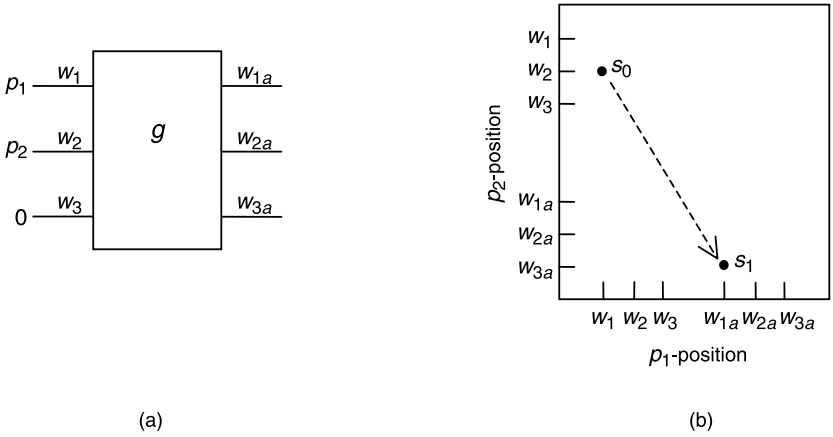


Figure 4.2
A state moves through configuration space. Particle p_1 moves from wire w_1 to w_{1a}, and p_2 moves from w_2 to w_{3a}.

lowing sections, we shall see how the configuration-space representation supports the introduction of quantumlike phenomena to our Fredkin-gate universes.

4.2.2 U2: A Universe with Noninterfering Superpositions

Suppose we modify the classical laws to allow a superposition of states to coexist. Rather than representing the state of the universe by a single point in configuration space, we assign a *weight* (between 0 and 1, inclusive) to each configuration-space point. The weights sum to 1. In U1, a single point changed its configuration-space coordinates at each unit-time interval. In U2, all weighted points move simultaneously, carrying their respective weights along. Each weighted point moves according to the same rules that governed the single point in U1.

We now say that each point in configuration space represents a *classical* state of the universe. The entire set of weight assignments in configuration space is a *quantum* state. In U2, the state of the universe is the quantum state, which we say is a superposition of its nonzero-weighted classical states. (When no ambiguity will result, I continue to speak of a "state," with "classical" or "quantum" left implicit.)

We may think of the weights in configuration space as probabilities. From that standpoint, the set of weight assignments specifies a probability

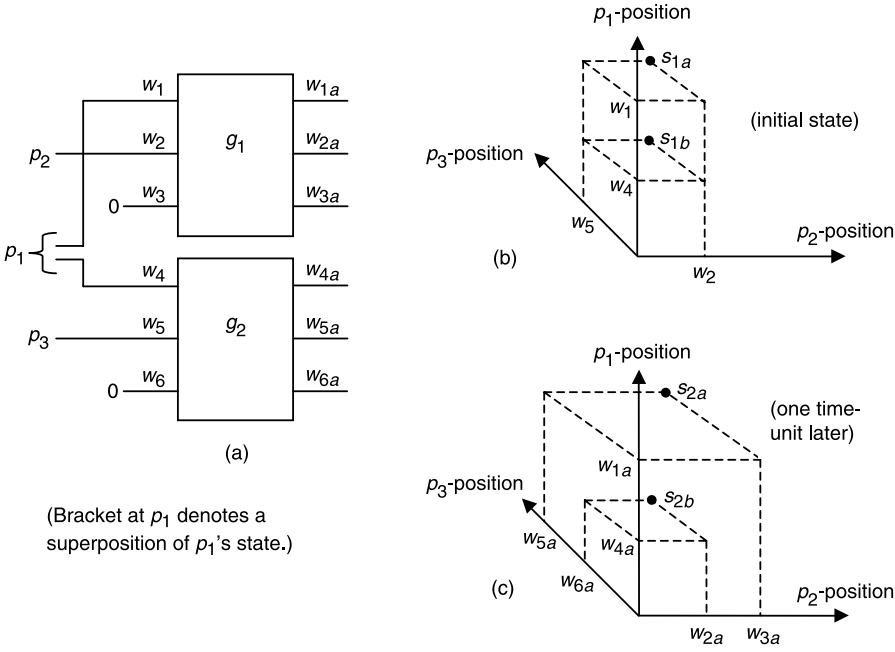


Figure 4.3
The configuration-space view: particles p_2 and p_3 observe particle p_1.

distribution as to what classical state the universe is in. In fact, though, the physical laws of U2 are not literally probabilistic; they are deterministic laws that push weights through configuration space.

Figure 4.3a shows a fragment of a Fredkin-gate circuit. (Here and throughout, unconnected wires are understood to connect to gates not shown.) Particle p_1 is in a superposition of two positions, w_1 and w_4. Particle p_2 is at w_2, and p_3 is at w_5. Figure 4.3b shows this situation from a three-dimensional cross section of configuration space, with one dimension for each of the three particles. (Each of the dimensions has a position for every wire in the circuit, but only a few of those positions are labeled in the figure.) States s_{1a} and s_{1b}, each with weight 0.5 (the weights are not shown), depict the superposed positions of p_1, but assign just one position to each of the other two particles.

Initially, the three particles' positions are mutually independent: in particular, p_2's position (like p_3's position) is the same whether p_1 is at w_1 or

w_4. One time unit later, though, the gates have correlated p_1 with p_2 and p_3 (as shown in fig. 4.3c). There is still a superposition of two world states, now s_{2a} and s_{2b}. In each state, p_2 and p_3 are consistent with where p_1 is in that state: if p_1 is now at w_{1a} (g_1's control-wire output), then p_2 has crossed over at g_1 (to w_{3a}), whereas p_3 has passed straight across at g_2 (to w_{5a}); but if p_1 is now at w_{4a} (g_2's control-wire output), then p_2 has passed straight across and p_3 has crossed over. Hence, the position of p_1 has been observed by p_2 and p_3. Although the universe still contains a superposition of two states for p_1, p_1's state *relative to* p_2's (to use Everett's terminology) is unambiguous: p_1 is at w_{1a} relative to p_2 at w_{3a}; p_1 is at w_{4a} relative to p_2 at w_{2a}. Similarly, p_1's state is unambiguous with respect to p_3's state.

Note the consistency of the two observations of p_1. There are only two possible outcomes: one state where p_2 crosses over and p_3 does not, so that only p_2 was diverted by p_1; and, symmetrically, a state where only p_3 was diverted by p_1. Hence, either state is consistent with p_1 being at w_1 or w_2, but not both. Moreover, it is easily verified that any subsequent observations—of p_1, p_2, or p_3—will maintain this consistency. By virtue of this consistent repeatability, the interactions with p_1 are what Everett calls *good observations*.

Prior to the observation, p_1 was in a superposition of two states. Subsequently, although this superposition continues, there are effectively two branches of the universe, each consistently and unambiguously showing one state of p_1. Thus, we might try to construe this interaction to model the apparent collapse of the quantum superposition—apparent, that is, from the standpoint of an observer embodied in U2.[5]

But that construal would be wrong. In fact, from within U2, there was never any apparent superposition to begin with. Hence, the observation did not appear to collapse any superposition. The problem is that there is no "interference"—no interaction at all—among the superposed classical states. Each such state has a unique immediate predecessor as well as a unique immediate successor (because, as is readily seen, a Fredkin gate's outputs uniquely specify what the inputs must have been, as well as vice

5. Here, in a leap of imagination, I envision an immense Fredkin-gate circuit that implements complex systems, including some that have advanced cognitive machinery like ours. Hence, that universe could embody intelligent observers.

versa). Thus, two superposed classical states never converge. Each evolves entirely independently of the other, moving through configuration space without interfering with the other.

Therefore, the superposition is evident only to an observer external to the entire universe who can examine configuration space directly. To any observer embodied in any ''branch'' of the universe (any element of the superposition), there is never any evidence of the existence of any other branch. Hence, this universe, as seen from within, appears entirely classical. It is indistinguishable from U1. In particular, the 1s behave like ordinary particles, just as in U1.

4.2.3 U3: A Quantish Artificial Universe

In this section, I present universe U3, with laws of physics that are analogous to real quantum mechanics under Everett's interpretation. Hence, I call U3 a *quantish-physics* model. This section largely recapitulates Everett's relative-state formulation of quantum mechanics, but with Fredkin gates substituted for quantum waves. The interference–observation duality of real-world physics—that superposed states interfere with one another if, and only if, no observation has distinguished among them—is a property of quantish physics as well.

The quantish-physics model extends and modifies the U2 model. U3's physics has three characteristics that distinguish it from U2 physics:

- multiple successor and predecessor states,
- complex rather than real-valued weights, and
- a binary-valued *sign* associated with each particle. (A particle's sign is analogous to spin, for example, in real quantum mechanics.)

In U2, each classical state has a unique successor and predecessor, so distinct states do not interfere. In the quantish U3 model, a classical state can have multiple immediate successors and predecessors. A configuration-space point's weight splits into components that each contribute to one of the point's immediate successors. The contributions of multiple predecessor-points to a common successor simply add.

To facilitate interference, quantish classical states are assigned complex weights rather than real-valued weights. The probability measure associated with a classical state is the squared magnitude of its weight. (It turns out, in

sec. 4.3.1, that this probability measure is something we can derive, rather than having to stipulate it.) In every quantum state, the classical states' probability measures sum to unity. When a classical state splits into two successors, its weight splits into two orthogonal components of the original weight (as elaborated below), so the sum of the successors' probability measures equals the predecessor's probability. When several configuration-space points contribute to a common successor point, the sum of the contributing weights has a squared magnitude that may be less than or greater than (or equal to) the sum of the contributing squared magnitudes. This possible inequality provides for destructive and constructive interference.[6]

Each quantish particle has a *sign*, whose value is either *plus* or *minus*. Each gate in a quantish-universe circuit has a *measurement angle*. A gate's measurement angle cannot change; like the circuit topology, it is simply built into the universe. But a particle's sign can change. Thus, each particle's sign is part of each quantish classical state and must be represented in quantish configuration space. Therefore, quantish configuration space has two dimensions for each particle: one, as in U2, for the particle's position, and the other for the particle's sign. Each sign dimension has just two discrete coordinate values, one corresponding to plus, the other to minus.

As with U2, quantish physics is defined by laws that say, for any classical state, where each particle next moves to. For quantish physics, the laws also say what the particle's next sign will be. As in the previous model, these laws translate into a rule that specifies the coordinates of a classical state's successor point in configuration space. The weight associated with the predecessor point moves to the new point.

But in the quantish model, a given particle in a given classical state can have two next positions, and two next signs, rather than just one of each. This multiplicity of destinations and signs corresponds to a fourfold split in the given classical state. That is, the given state has four successor states rather than a single successor. There is one successor state for each of the four permutations of destination and sign for the given particle. Thus, no successor state shows the particle simultaneously at more than one position, or with more than one sign. Rather, there is a distinct classical state for each of the alternatives.

6. Particles' trajectories through configuration space correspond to the lines in Feynman diagrams (Feynman 1985).

The given state's weight divides among the four successors, as described below. More generally, in a given classical state, there may be n particles with two next positions and signs each. Then, there are 4^n successor states, one for each combination of the binary next-position and next-sign choices for each of the n particles.

Defining quantish physics requires specifying:

- a rule by which a particle's weight divides between the particle's next positions and signs; this rule says how a particle moves through a gate; and
- the rule by which weights combine when multiple predecessors have one or more successor points in common.

How particles move through gates is explained just below. The rule for combining weights is trivial: as mentioned above, when several configuration-space points each contribute a portion of their weight to a common successor-point, the contributed weights simply add. This, together with the fact that a classical state's successors are a function of that state alone (regardless of any other classical states superposed in the quantum state), ensures that the quantish-physics state-succession (like real-world quantum-state evolution) is linear. That is,

$$\text{successor}\,(q_1) + \text{successor}\,(q_2) = \text{successor}\,(q_1 + q_2),$$

where the successor function maps a quantum state onto its successor quantum state. States q_1 and q_2 are quantum states, and $q_1 + q_2$ is the quantum state whose weight at each configuration-space point is the sum of q_1 and q_2's weights at that point.

A particle at the control-wire input to a quantish gate simply passes through to the control-wire output, as in U1 and U2. Its sign remains the same. However, a particle that is at a gate's switch-wire in a given classical state behaves differently than in U1 and U2: roughly speaking, the particle emerges at *both* of the gate's switch-wire outputs (as suggested by the bracket notation beside the gate in fig. 4.4a, depicting a superposition of positions for particle p_1), and with both signs at each destination (as suggested by the $\pm$ notation). More precisely, as mentioned above, the different destinations and signs occupy four distinct successor states.

The gate's measurement angle is Q, as depicted in the figure. Let us say that the gate *measures* particle p_1 with respect to measurement angle Q. The original weight c_1 splits among the just-mentioned successor states, as follows:

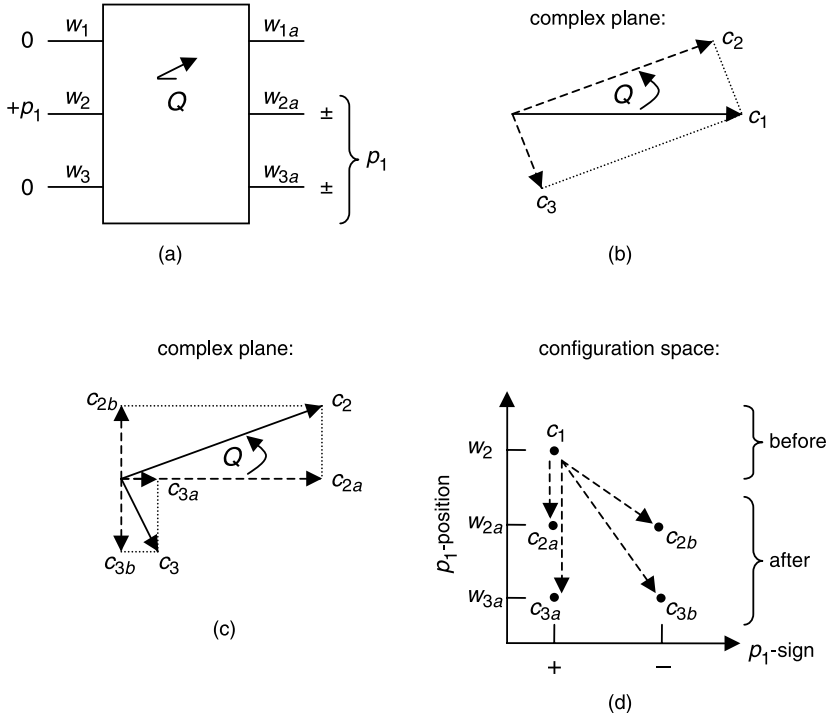


Figure 4.4
A classical state with weight c_1 splits into four successors.

▪ First, we define a *measurement vector* in the complex plane. If, as in figure 4.4, the switch-wire particle's sign is plus, then the measurement vector is the weight c_1 rotated in the complex plane by the gate's measurement angle Q (fig. 4.4b). If instead the switch-wire particle is minus, the measurement vector is the weight rotated by $Q - \pi/2$—that is, it rotates by Q, and then back by 90 degrees. (The rationale for this orthogonal twist will become apparent in the following section.)

▪ The weight c_1 divides into two orthogonal components, c_2 and c_3. One component is parallel, the other perpendicular, to the measurement vector in the complex plane (fig. 4.4b again). Call these the measurement-parallel and measurement-perpendicular components of the original weight. For a plus-state particle, the two components are thus rotated by Q and $Q - \pi/2$, respectively, from c_1. Their magnitudes are $|c_1|\cos(Q)$ and $|c_1|\sin(Q)$, respectively.

- If, as in figure 4.4, the classical state has no particle at the gate's control wire, the measurement-parallel component of the state's weight moves to a successor state (or rather, to a pair of states, subdividing again as specified just below) in which the switch-wire particle passes straight across. The measurement-perpendicular component similarly moves to states in which the switch-wire particle crosses over. If instead a control-wire particle is present, the opposite correspondence holds: the measurement-parallel component corresponds to crossing over, the measurement-perpendicular component to passing straight across.
- The measurement-parallel and measurement-perpendicular components each subdivide further into two components, one parallel, one perpendicular to the original weight c_1 (fig. 4.4c). The parallel components (c_{2a}, c_{3a}) move to the successors in which the switch-wire particle has the same sign it had in the original classical state. The perpendicular components (c_{2b}, c_{3b}) move to the other successors, in which the particle's sign has changed (fig. 4.4d).

Thus, the weight-splitting rule twice decomposes a weight into a pair of orthogonal components.[7] The sum of the components therefore equals the original weight: $c_1 = c_2 + c_3 = (c_{2a} + c_{2b}) + (c_{3a} + c_{3b})$. Also, at both steps, the probability measure, defined as a weight's squared magnitude, is conserved, according to the Pythagorean theorem: $c_1^2 = c_2^2 + c_3^2 = (c_{2a}^2 + c_{2b}^2) + (c_{3a}^2 + c_{3b}^2)$. Finally, note that in the special case where the measurement angle is zero, the above rule (applied to a plus-sign particle) is equivalent to U1 or U2 state succession—since the measurement-orthogonal component is zero, no state-splitting occurs, and the particle entirely passes straight across, or entirely crosses over, depending on whether a control-wire particle is present. (The next section shows that a measurement angle of zero is not privileged in this respect, however. Rather, any measurement angle can fail to produce state-splitting under certain circumstances.)

7. Decomposing quantum weights into orthogonal components, and swapping some of those components, is a bit reminiscent of the velocity-vector decomposition for Newtonian collisions discussed above in section 3.2. However, the Newtonian decomposition was in physical space, whereas here there is a decomposition of complex weights in the complex plane. Positions and directions in the complex plane have no correspondence to positions and directions in physical space or in configuration space.

The above description specifies the fourfold split of a classical state for a single switch-wire particle in that classical state. When a classical state has n particles at switch wires, there are 4^n successor states, as noted above. The n four-way splits are applied in succession, in any order.[8] (As the reader may verify, for a given switch-wire particle, each of the four split-apart successor weights equals the original weight multiplied by a complex factor. Since such multiplication is commutative and associative, one may think of the n splits as occurring in any order, or simultaneously.) This 4^n-fold splitting also conserves both probability and quantum weight, for it is equivalent to n successive fourfold splits, each conserving probability and quantum weight.

4.2.4 Successive Measurements in U3

The laws of U3's quantish physics are now completely specified.[9] A brief look at the effects of passing a particle through the switch-wire inputs of successive gates will elucidate important aspects of these laws, in preparation for examining their quantumlike properties.

Figure 4.5 extends figure 4.4a. In figure 4.5, gate g_1's upper switch-wire output connects to g_2's upper switch-wire input. (The gate's other switch-wire output diverts to some other gate, not shown.) Gate g_2 has the same measurement angle Q as g_1. The arrow at wire w_2 designates the weight c_1 (from fig. 4.4b) associated with the state in which p_1 is at that wire, with plus sign. (For convenience, the initial weight has angle zero in each of the scenarios to follow. But there is nothing privileged about that orientation. The quantish rules decompose a quantum weight according to a measurement vector whose orientation is defined *relative to* the original weight. Hence, rotating the original weight would do nothing more than to correspondingly rotate all the subsequent ones too.)

The arrow at wire w_{2a} designates the measurement-parallel component-weight c_2 (fig. 4.4b again), which is divided between the two successor

8. If a gate has particles at both switch-wire inputs, this formulation allows some successor states that have two particles at the same position. However, that does not occur in any of the examples here.

9. For a given quantish universe, we must of course also specify the circuit topology and an initial state. In keeping with the simplicity criterion discussed in the previous chapter, we might choose an initial state that assigns all the quantum amplitude to some single classical state.

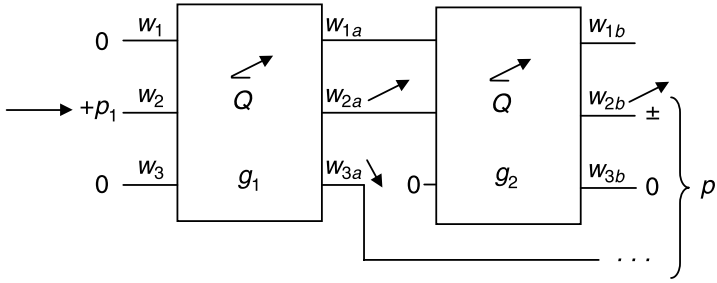


Figure 4.5
Using the same measurement angle twice in a row causes no further state-splitting.

states in which p_1 reaches w_{2a} (one successor with plus sign, the other with minus sign); analogously for the arrows at w_{3a} and w_{2b}. It turns out, as explained just below, that because the second gate has the same measurement angle as the first, it causes no further state-splitting—that is, p_1 proceeds straight across to w_{2b}, and with no change in the weights associated with the states that assign the particle's superposed signs. In particular, p_1 never emerges from wire w_{3b}.

In fact, the two states (with weights c_{2a} and c_{2b}, as in fig. 4.4c) in which p_1 reaches w_{2a} each have, at that point, the usual four successor states (after p_1 passes through the next gate, g_2), one for each combination of p_1's next position and sign. But both states have the *same* four successors, so each of those successors receives a component of c_{2a} and also of c_{2b}. In each of the two successor states in which p_1 crosses over to w_{3b}, the two components sum to zero. In the other two successors, the components sum to recreate c_{2a} and c_{2b}. That this happens can be verified by applying the state-splitting rule in detail. But there is also a more intuitive explanation:

▪ The second gate, g_2, with the same measurement angle Q as g_1, turns out to decompose each of the weights c_{2a} and c_{2b} with respect to the same measurement vector that was used for g_1's decomposition of c_1 into c_{2a} and c_{2b}. This is so for c_{2a} because c_{2a} is parallel to c_1, so rotating c_{2a} by Q yields the same measurement vector as rotating c_1 by Q. On the other hand, c_{2b} is perpendicular to c_1—but p_1 has become minus in the state whose weight is c_{2b}, and so the rule that adds an orthogonal twist to the measurement vec-

tor for a minus particle now cancels c_{2b}'s perpendicularity, so that its Q-rotated measurement vector is also parallel to c_1's.

- Once the components have been decomposed with respect to the same measurement vector as before, the measurement-parallel component moves to configuration-space points that have p_1 passing straight across, the measurement-orthogonal component to points that have p_1 crossing over. But the measurement-orthogonal component is zero, because that component of c_1 was diverted away at the first gate. Thus, the measurement-parallel component simply remains intact at the second gate.
- Finally, the states that distinguish p_1's signs keep the same respective weights. The state-splitting rule either leaves a particle's sign unchanged, as well as the orientation of the corresponding weight, or else complements the particle's sign, along with making an orthogonal twist to the corresponding weight. A sequence of two such complements and twists both restores the original sign, and reestablishes parallelism with the original weight. Thus, all resulting weights are either parallel to the original and assign the same sign to the particle, or are perpendicular and assign the opposite sign. Therefore, the reconstructed measurement-parallel weight c_1 must decompose into the same components as before, respectively assigning the same signs to p_1.

Thus, after the particle passes through the first gate, the decomposition of the state's quantum weight into orthogonal components is such that the particle's passage through the second gate causes no further state-splitting. Although the usual state-splitting rule still applies at the second gate, the split-apart and reconverged weight components combine and cancel to reconstruct the weights exactly as they were just before the particle's passage through the second gate. Hence, the particle emerges only from the upper switch-wire there.

Recall, though, that each quantum weight is associated not with an individual particle, but rather with a point in configuration space—and hence, with a configuration of *all* particles. The analysis just given proceeded as though there were no other particles simultaneously passing through switch wires elsewhere in the circuit; hence, in that analysis, only p_1's passage through g_1 splits the global quantum weight. But what if other particles elsewhere in the circuit also induce such splits simultaneously? Does

their effect on the global weights' orientations destroy their alignment with g_2's measurement angle and thus prevent that alignment from preserving p_1's exclusive position at the upper switch-wire?

There turns out to be no such problem. As already noted, each split component of the quantum weight is the multiplication of the original weight by some complex factor (specified by a gate's measurement angle); and such multiplication is commutative and associative, so a series of such multiplications can be performed in any order and will still yield the same result. In other words, the simultaneous state-splits can be thought of instead as occurring sequentially, in any order. In particular, then, let us suppose that when p_1 passes through g_1, the corresponding state-splitting occurs after any other simultaneous splits; and suppose that when p_1 next passes through g_2, the corresponding state-splitting occurs *before* any other simultaneous splits. So we can just apply the above analysis independently to each of the already simultaneously split weights at the point when p_1 passes through g_1. The conclusion, then, is that for each such weight, no further state-splitting (after reconvergence and cancellation) is induced by p_1's passage through g_2. Thus, we can safely consider the state-splitting at g_1 and g_2 independently of any simultaneous splits induced elsewhere in the circuit.

In the alternative circuit of figure 4.6, g_1's lower switch wire, rather than its upper one, connects to g_2. By reasoning similar to the above, g_2 again causes no further state-splitting. In this case, it is only the measurement-orthogonal component of the original weight that reaches the configuration-space points that correspond to p_1 reaching g_2. Rather than entirely passing straight across, here p_1 entirely crosses over, arriving back at the top at switch wire w_{2b}.

Combining the results from figures 4.5 and 4.6, figure 4.7 shows the result of connecting *both* of g_1's switch-wire outputs to the corresponding inputs of g_2. Since quantish state-succession is linear, the weights reaching the four states in which p_1 emerges from g_2 are the sums of those weights in the previous two examples. The measurement-parallel component of c_1 follows the states that have p_1 passing straight across the upper switch-path at both gates. And the measurement-perpendicular component follows p_1 crossing over at the first gate, then back again at the second, thus also arriv-

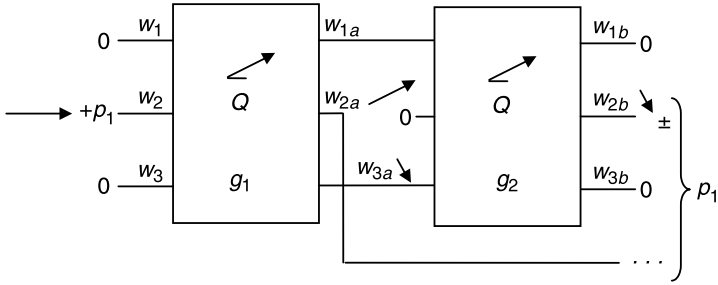


Figure 4.6
Again, reusing the same measurement angle causes no further state-splitting.

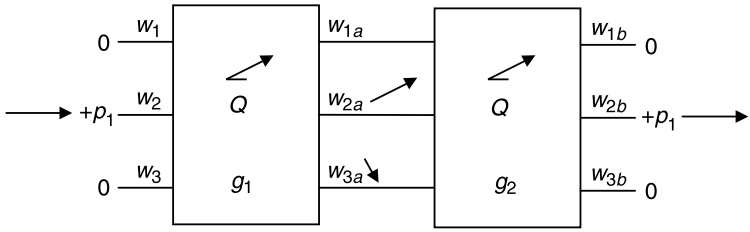


Figure 4.7
Connecting *both* switch wires to a gate with the same measurement-angle undoes the first gate's splitting.

ing at g_2's upper switch-wire output. The two components sum there to recreate the original weight (and with p_1 restored exclusively to its original sign).

Finally, figure 4.8 illustrates the effect of a succession of different measurement angles. For the states in which p_1 appears at g_2's upper switch wire, g_2 divides the corresponding weights into measurement-parallel and measurement-perpendicular components, but with respect to a different measurement vector than at g_1 (in fig. 4.8, the weight corresponding to the arrow at w_{2a} divides into the orthogonal components shown at w_{2b} and w_{3b}). The measurement vector at the second gate differs from that at the first by Q_2–Q_1. Hence, at the second gate, the measurement-parallel and measurement-perpendicular components have squared magnitudes of $\cos^2(Q_2–Q_1)$ and $\sin^2(Q_2–Q_1)$, respectively, times the squared magnitude of the original weight. Figure 4.5 is the special case in which $Q_1 = Q_2$.

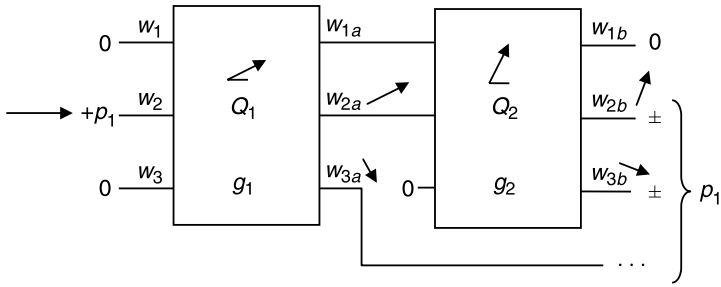


Figure 4.8
Using a succession of different measurement angles makes states split.

4.3 Quantumlike Properties of Quantish Physics

The laws of quantish physics, like the laws of U1 and U2, are *local*: only nearby things interact. The destinations (and new signs) of a particle at a switch-wire of some gate in some superposed classical state depend only on the particle's current sign, the gate's measurement angle, and whether there is a particle at the control wire of the same gate in the same classical state. Similarly, the destination and sign of a control-wire particle at some gate in some state depend only on that gate and that particle in that state.

Thus, there is no action at a distance with respect to circuit-topology space, or with respect to configuration space. And, of course, the quantish-physics laws are entirely deterministic. I now demonstrate that these local, deterministic laws support phenomena like those of the real quantum world: seeming indeterminacy of quantum states, interference of superposed outcomes, and interference–observation duality.

4.3.1 Apparently Nondeterministic Outcomes and the Uncertainty Principle

In figure 4.9, particle p_1 "splits" at g_1 (as in fig. 4.4a), and is then observed at gates g_2 and g_3 (as in fig. 4.3 back in U2). (Here and throughout, when I show gates wired in series, seemingly disconnected inputs shown at gates later in the series are assumed to be synchronized, by circuitry not shown, to arrive there simultaneously with inputs from earlier in the series. Thus, in fig. 4.9, p_2 and p_3 arrive at gates g_2 and g_3 simultaneously with p_1.)

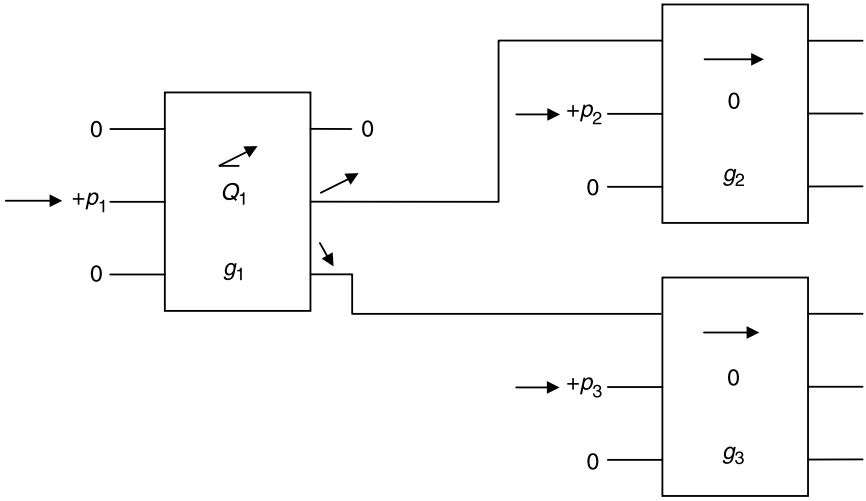


Figure 4.9
Particles p_2 and p_3 observe p_1's position.

Assume that p_2 has already just passed through a gate with the same measurement angle (namely, 0) as g_2's, and similarly for p_3. Now, in the successor states that have p_1 arriving at g_2's control wire, p_2 entirely crosses over. In those same states, particle p_3 entirely passes straight across, since the states in which p_1 arrives at g_2's control wire do not have p_1 arriving at g_3's. Similarly, in states in which p_1 does arrive at g_3, p_3 crosses over and p_2 passes straight across. Thus, as in figure 4.3, the two observations are consistent: p_1 is always observed at exactly one of its two possible destinations.

From within the quantish universe, then, it appears that p_1 arrives at one gate or the other, but never both; every successor state is consistent with there being just one destination. Although different successors with different destinations remain in superposition, they have no effect on one another (unless they later reconverge in configuration space, as addressed in the next section). However, because g_1's measurement angle is oblique (i.e., neither horizontal nor vertical), the particle's destination cannot be uniquely specified in advance—because, in reality, it will have both destinations, notwithstanding appearances to the contrary from the point of view of any individual superposed classical state in the quantish universe.

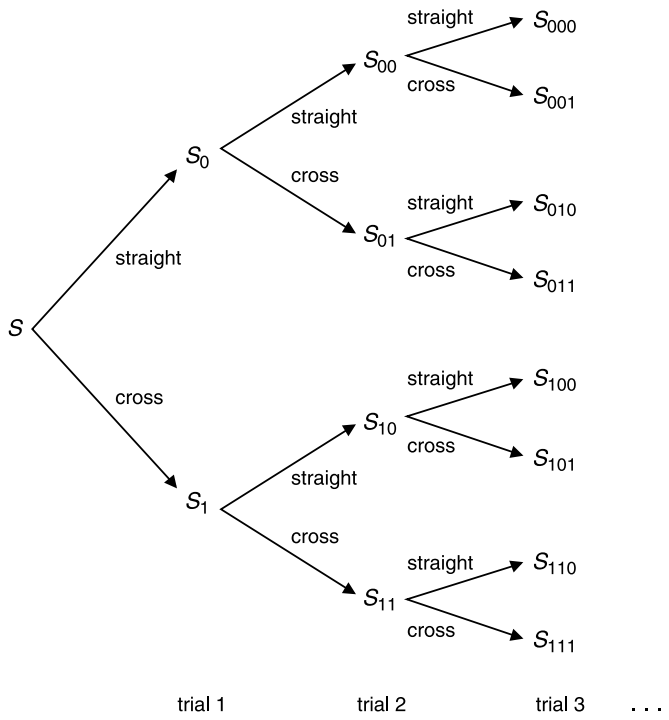


Figure 4.10
On each next trial, each state has a pair of successors corresponding to the two superposed outcomes.

Suppose observers embodied in the quantish universe conduct a number of trials with an apparatus such as in figure 4.9, recording the result of each trial. The first trial transforms a given state into a pair of successor states, one for each outcome (p_1 passed straight across, or crossed over). Each successor state in turn leads to two successors for each next trial, forming a tree of states as shown in figure 4.10. After n trials, there are 2^n successor states. In each, the cumulative record shows the sequence of outcomes leading to that state. For each possible permutation of outcomes in the sequence of trials, there is some superposed state in which that sequence occurred, and in which that sequence is reflected in the cumulative record in that state.

There is one eventual state in which all the recorded outcomes showed p_1 passing straight across, and another in which p_1 always crossed over. In each of the (exponentially many) other superposed states, though, the cu-

mulative record shows examples of passing straight across and examples of crossing over. By virtue of such cumulative records, the outcome of the trials thus seems—in almost all superposed states within the quantish universe—to be nondeterministic, coming out one way at times, another way at other times.

That, of course, is how an impression of quantum indeterminacy arises in the real universe as well: each individual observation is definite, but the recorded distribution of outcomes in identical repetitions of the experiment looks random. Moreover, the apparent nondeterminism is quantifiable. Given enough trials, almost all the weight in configuration space will be assigned to states whose cumulative records show that p_1 passed straight across in approximately $\cos^2 Q$ of the trials, and crossed over in approximately $\sin^2 Q$, of the trials.

That distribution occurs because each successive squared-amplitude split into the sine-squared and cosine-squared components is mathematically the same as multiplying each next probability by those respective components. And the law of large numbers assures that the outcome distribution over many identical-probability trials almost always approximates the probability distribution that applies to each of the trials. Thus, for example, if you flip a fair coin 1,000 times, you'll almost certainly see heads roughly 500 times. That's because if you could enumerate all $2^{1,000}$ (equally probable) possible sequences of 1,000 heads or tails (don't try it at home—$2^{1,000}$ is far more than the number of atoms in the universe), you'd find that almost all those sequences have a roughly 50–50 distribution of heads and tails. (There are, of course, a great many sequences among the $2^{1,000}$ that instead are highly skewed. But there are exponentially many more that are roughly even, making the skewed ones vanishingly rare by comparison.)

Think of the quantum weights as being the ultimate stuff of the quantum universe; quantum weight flows through configuration space, assigning amplitudes to various classical world-states. For almost all of this universe stuff, the corresponding classical state's cumulative record of a sequence of many trials of our experiment shows a distribution in which the particle has passed straight across, or crossed over, approximately $\cos^2 Q$ or $\sin^2 Q$ of the time, respectively. Those, then, are the apparent probabilities of the two outcomes, as seen from almost everywhere within the quantum universe; given sufficiently many trials, large deviations from that distribution

occur in an exquisitely rare portion of the quantum weight. Hence, it is unsurprising that we, in particular, observe a distribution that approximately reflects the square-magnitude split of the quantum weights, whenever we perform a quantum experiment with a large number of trials—it would be unexpected, to say the least, to find ourselves in the rare portion of quantum weight that shows a very different distribution of outcomes. (This argument for quantifying apparent nondeterminism by appeal to cumulative records in configuration space is adapted directly from Everett.)

Notice, though, that a minuscule portion of the quantum weight is not necessarily the same as a minuscule portion of the associated configuration space. For instance, suppose we conduct a thousand trials of a quantum observation for which outcome *A* gets 75 percent of the quantum squared amplitude, and outcome *B* gets 25 percent. If we trace the flow of quantum weight from the start of those trials, we see it divides among $2^{1,000}$ branches—but it does not divide equally. In almost all of those branches, approximately half of the cumulatively recorded outcomes are *A*, and half *B*.

- But in the branches associated with almost all of the quantum *weight*, the recorded distribution is approximately 75 percent *A*, 25 percent *B*. As just noted, then, the roughly 3:1 distribution that we (almost always) observe after many such trials is unsurprising if we think of quantum weight as the stuff of the universe.
- Less metaphysically, Deutsch (1999) and Wallace (2003b) have pointed out that we can reach the same conclusion—that apparent probability corresponds to quantum amplitude squared—by appeal to decision theory. First, notice that any distribution of quantum outcomes can be implemented (to an arbitrarily good approximation) by an underlying quantum process involving outcomes of equal magnitude. For example, a 3:1 distribution can be implemented by an experiment having four equal-magnitude outcomes, with three of them labeled *A* and one labeled *B*. Given many identical trials each involving equal-magnitude outcomes, the expected approximate cumulative distribution occurs in almost all quantum branches *and* in almost all of the quantum weight. For example, given the equal four-way split just proposed, almost all branches—and the branches associated with almost all of the quantum weight—exhibit a

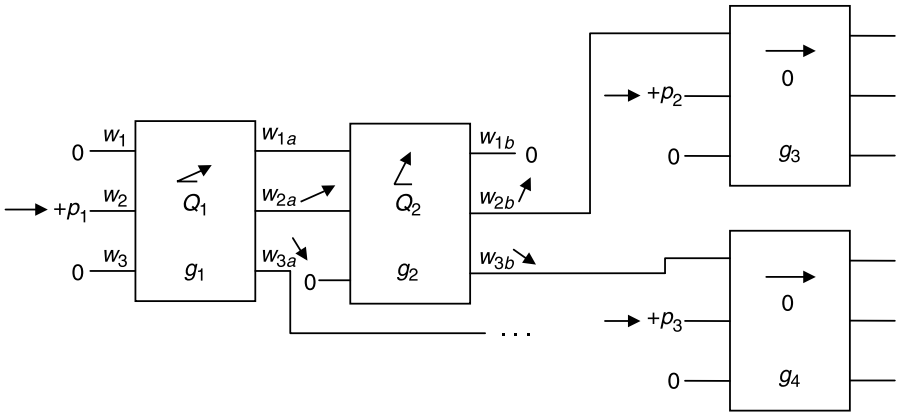


Figure 4.11
An observation distinguishes between two outcomes of passing p_1 through a succession of gates with distinct measurement angles.

roughly even cumulative distribution among the four outcomes, about 75 percent of which are labeled *A*.

By symmetry, a rational agent has no more reason to bet on one of *n* equal-magnitude outcomes than on any other; from a decision-theoretic standpoint, then, those outcomes should be deemed equally probable. If, further, we have no reason to care how a given quantum experiment is implemented (provided that it has the specified amplitude distribution of outcomes), then we can model any such experiment as being implemented by suitably labeled underlying equal-magnitude outcomes, giving us access to the symmetry argument. Effective probability then will always turn out to correspond to the square of quantum amplitude.

Figure 4.11 extends figure 4.8, observing (as in fig. 4.9) whether p_1 emerges at w_{2b} or w_{3b}. Over many such trials (counting only those occasions on which p_1 passes through g_2 at all), the typical cumulative record would show p_1 emerging at w_{2b} with frequency $\cos^2(Q_2–Q_1)$, and from w_{3b} with frequency $\sin^2(Q_2–Q_1)$.

Thus, in particular, if both gates have the same measurement angle, p_1 will always be observed to emerge at g_2's upper switch wire (as in fig. 4.5). We may therefore say that, having passed through g_1, p_1 now has a definite state with respect to g_1's measurement angle Q_1—meaning that there is no apparent nondeterminism (that is, no multiplicity of

superposed outcomes) regarding p_1's next destination if p_1 runs through another gate with that same measurement angle.

But having a definite state with respect to one measurement angle always means having an indefinite state with respect to all angles oblique to that one (i.e., neither parallel nor perpendicular). Quantish configuration space does not separately encode (i.e., does not provide a distinct configuration-space dimension for) a particle's state with respect to each possible measurement angle. Rather, configuration space designates just a single binary attribute for each particle, its sign. That attribute corresponds to a definite state with respect to some measurement angles ($\pi/2$ and its multiples), but not with respect to other angles.

But for any other measurement angle, there is some possible superposition of signs that creates a definite state with respect to that angle (and with respect to angles parallel or orthogonal to that angle), but not with respect to angles oblique to that angle. As just noted, figure 4.11 illustrates this principle if $Q_1 = Q_2$. Thus, as seen from within the quantish universe, there is nothing privileged about being in a definite state with respect to the angles 0, $\pi/2$, and so forth. True, a definite state with regard to those angles corresponds to a pure plus or minus, whereas a definite state with regard to other angles corresponds to some superposition of plus and minus. But that distinction is undetectable from within the quantish universe; experiments there simply show, in either case, a definite state with respect to some measurement angles, but not with respect to others. (Physicists would say that the plus and minus states correspond to an arbitrary *basis* in configuration space.)

Thus, a particle's inclination to cross over or not at the next gate cannot be made definite with respect to all possible measurement angles at the next gate. Eliminating apparent nondeterminism with respect to a given measurement angle (by observing a particle's inclination to cross over with respect to that angle at a previous gate) necessarily creates apparent nondeterminism with respect to all angles oblique to that previous angle (the particle will both cross over and pass straight across if an oblique angle is used next; neither of those outcomes will be nullified by the interference of reconverging superposed states).

A consequence of this inherent trade-off (between definiteness with respect to one angle and definiteness with respect to another) is the false-

hood of so-called *hidden-variable* accounts of apparent quantum nondeterminism. In the quantish world, a hidden-variable account would propose that each particle has some internal attributes that are not directly observable; these attributes give the particle a definite state with respect to every angle, though the attributes' hiddenness makes the states seem random. The foregoing analysis, though, shows that hidden-variable theories are wrong. (Sec. 4.3.4 below presents a quantish analog of the quantum EPR experiment, which provides an especially dramatic disproof of hidden-variable theories.)

The trade-off between definiteness with respect to one angle and definiteness with respect to another recapitulates Heisenberg's uncertainty principle in the quantish universe. Heisenberg would say that we cannot simultaneously know a quantish particle's state with respect to two mutually oblique measurement angles. But contrary to Heisenberg's own interpretation of his uncertainty principle, the principle does not address a limit on what we can know about a particle's state—hence, the uncertainty principle is not about uncertainty! Rather, it is a statement about the deterministic, fully knowable behavior of superposed states and their interactions. It is a straightforward consequence of the lack of separate configuration-space dimensions to encode a particle's state with respect to mutually oblique measurement angles.

Similarly, in real-world quantum physics, configuration space does not separately encode a particle's position and momentum. Only position is encoded (or, equivalently, only momentum, or only some particular combination of the two). The basic physical law of motion says that an undisturbed particle spreads in all directions at light speed—or rather, that a weight in configuration space spreads at light speed (with no change in phase) into a filled-in sphere across the three configuration-space dimensions that specify the particle's position. The particle thus has a maximal superposition of momenta.

But the spread can be confined to a smaller envelope by arranging a superposition of appropriately phased weights for the particle's positions. The weights assign a superposition of positions to the particle, but interference among them constrains their spread, thus limiting the superposition of the particle's momenta. In the limit, a superposition of *all* positions can produce a single, definite momentum.

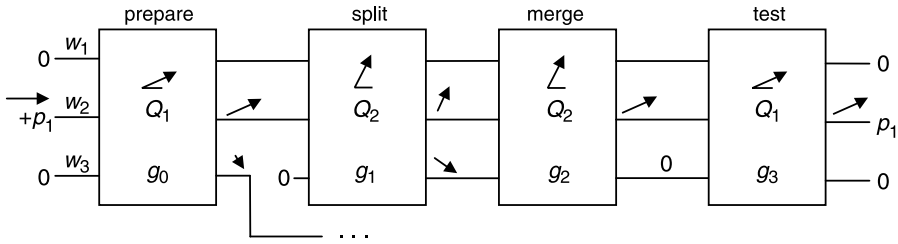


Figure 4.12
Superposed states remerge and interfere.

The sharing of a single configuration-space dimension for a given particle's position and momentum along a given spatial dimension thus creates an ineliminable trade-off between definiteness of positions and definiteness of momenta. In the quantish analog, the sharing of a single configuration-space dimension for a particle's sign creates an ineliminable trade-off between definiteness with respect to one measurement angle and definiteness with respect to any angle oblique to that angle.

4.3.2 Interference of Superposed States

Having seemingly nondeterministic outcomes is a step toward having quantumlike phenomena. Beyond mere nondeterminism, though, we still need evidence of the fundamental quantum duality, which requires superposed states that can mutually interfere, unless distinguished from one another by observation. I now demonstrate such behavior in the quantish-physics model.

In figure 4.12, the first gate, g_0, prepares particle p_1 by putting p_1 in a definite state with respect to Q_1 before sending p_1 to g_1. Particle p_1 also diverts from g_1 (at the lower switch-wire output), but the discussion that follows only addresses the case in which it reaches g_1. The states in which p_1 diverges do not interfere with the states under discussion, since they are separated along the p_1-position dimension.

At g_1, p_1 splits, using measurement angle Q_2. Then, at g_2, p_1 remerges (as in fig. 4.7), reconstructing the weight with which p_1 entered g_1, thereby reestablishing p_1's definite state with respect to Q_1. Finally, g_3 performs a test to verify that p_1 has a definite state with respect to Q_1. Suppose p_1 were then observed emerging from g_3 (that observation is not shown here,

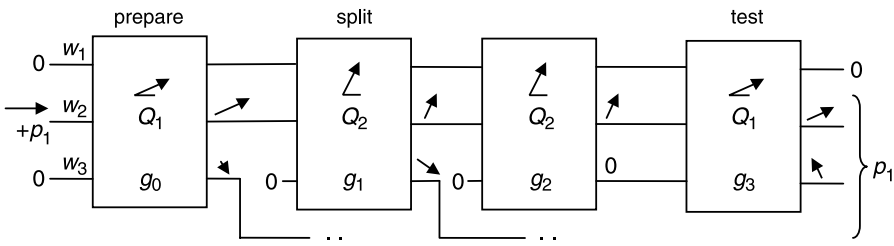


Figure 4.13
Here, one path diverts to prevent merging at g_2.

but would be similar to the observation of p_1's emergence from g_2 in fig. 4.11). Over many such trials, particle p_1 would always be observed to arrive at g_3's upper switch-wire output.

In figure 4.13, one path to g_2 is disconnected (as in fig. 4.5), circumventing the merging. In those states in which p_1 does reach g_2, p_1 already has a definite state with respect to Q_2, so p_1 entirely passes straight across and keeps its definite state with respect to Q_2. Thus, p_1 does not have a definite state with respect to Q_1, so—unlike in figure 4.12—p_1 is split by g_3, emerging both at the upper and lower switch-wires.

We are now in a position to see the effects of superposed states' interference in the quantish-physics model. Contrasting figures 4.12 and 4.13, we see a genuine quantum-interference phenomenon: figure 4.12, compared to figure 4.13, provides an additional path by which p_1 might reach g_3's lower switch-wire output. Yet p_1 emerges there *less* often (in fact, never) with the extra path provided than without that path. This contrast is inexplicable on the classical assumption—which otherwise seems correct, from within the quantish universe, as seen in the previous section—that p_1 is a particle-like entity that exists at just one wire at a time.

Only by acknowledging the simultaneous reality of p_1's superposed positions at both of g_1's switch-wire outputs can we (or any observer embodied in the quantish universe) account for the possibility that those states can interfere with one another when a path is provided to convey the interfering influence. The interference is achieved, of course, by the summation of complex weights at common successor states, as discussed in section 4.2.4; opposite weights cancel when added. In figure 4.13, diverting p_1 from reconverging to the same position thereby diverts the corresponding

configuration-space path from reconverging, thus circumventing the interference that the reconvergence would bring.

The setup of figures 4.12 and 4.13 is analogous to the real-world double-slit experiment, in which a particle is in a superposition of states (passing through slit 1, or slit 2):[10]

- *Destructive* interference among the superposed states reduces the likelihood of the particle's arrival at certain points along the backdrop. But blocking one of the two possible paths thereby blocks that interference, returning the probability to normal. (The two slits are like the two switch-wire inputs to g_2 in figs. 4.12 and 4.13. The diversion away from the lower input to g_2 in fig. 4.13 is like blocking one of the two slits.)
- Less dramatically paradoxically, *constructive* interference increases the probability of arrival at certain points, so that the probability exceeds what the sum of the two single-slit curves would predict. Correspondingly, the frequency of arrival at g_3's upper switch-wire output when both paths are provided is greater than the sum of the frequencies when just one path or the other is provided.

4.3.3 Blocking Interference via Observation

If inhabitants of a quantish-physics universe perform the above experiments, they face the same apparent paradox that physicists in the real universe encounter. When a "split" particle is observed as in figure 4.9, the results consistently and unambiguously show that the particle reached one destination or the other, but not both. Yet, comparing the behavior of the circuit in figure 4.12 with that in figure 4.13, there is a demonstrable interference effect that is explicable only on the assumption that the particle does reach both destinations (which is indeed the case, as we privileged observers of configuration space can see, looking in from outside the quantish universe).

Let us sharpen the paradox further. Suppose inhabitants of the quantish universe try to *observe* p_1 on its way to g_2—that is, after p_1's path splits and

10. We might also construe this setup as an analog of the Stern–Gerlach experiment (see, e.g., Cohen-Tannoudji, Diu, and Laloë 1977). Particle p_1's sign is analogous to a real-world particle's spin; g_1 and g_2 together correspond to a Stern–Gerlach module that diverges and then reconverges particles' paths according to their spin with respect to a certain axis (analogous to the gates' common measurement angle).

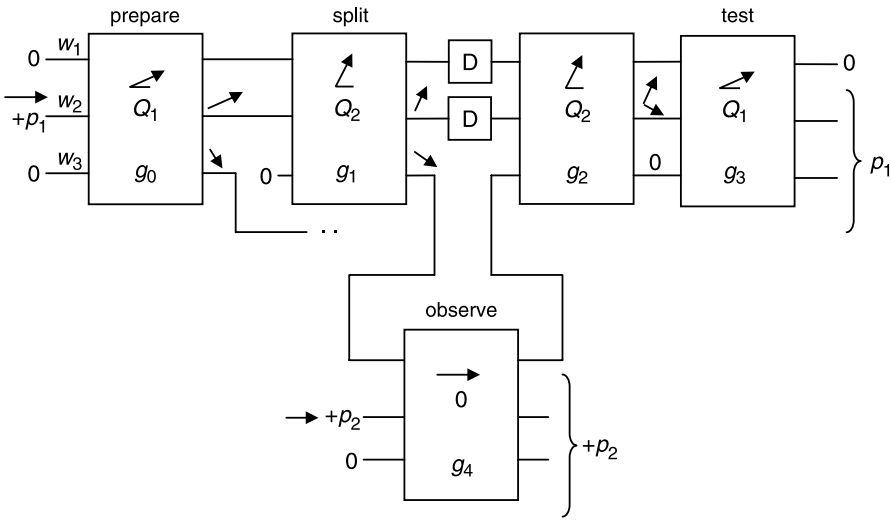


Figure 4.14
An observation circumvents subsequent interference.

before the path remerges. Figure 4.14 shows a setup in which p_2, at gate p_4, crosses over or not depending on whether p_1 passes through g_4. Particle p_1 is then routed into g_2 as before. Delay gates (labeled D) have been inserted at the other two paths to g_2 to maintain synchronization. (A delay gate is an ordinary Fredkin gate. The wire shown is its control wire. The switch-wire inputs, not shown, have no particles present.)

Particle p_1 is unaltered by the observation. Classically, then, the observation should not change the outcome of the experiment. But in quantum physics, making an observation to distinguish two superposed states blocks any subsequent interference between those states—and that is just what happens here. We find the same bizarre result as in the real universe when we observe which slit the electron came through: the interference disappears. Without the interference, p_1 can once again emerge from either of g_3's switch-wire outputs.

The configuration-space explanation of this phenomenon is straightforward. Although p_1 reconverges to the upper switch-wire after passing through g_2, occasioning a reconvergence along the corresponding weights' p_1-position dimension in configuration space, the weights remain separated along their p_2-position dimension, because p_2 does not reconverge.

Since the weights thus fail to reconverge to the same configuration-space point, they do not add together and interfere. (The two vectors shown at g_2's upper switch-wire output represent the superposed weights separated along the p_2-position dimension. If instead those weights had reconverged and added, they would have reconstructed the weight shown at g_0's upper switch-wire output.) The outcome, as seen from any of the successor states, is just as though p_1 had traversed just one path or the other through g_1 (as the classical view would have it), but not both.

Even a so-called *negative observation* results in the absence of interference. In the states in which p_1 does not reach g_4, p_1 does not interact with p_2. But that very absence of interaction—that is, a negative observation of p_1 by p_2—is fully informative about p_1's whereabouts: if p_2 does not cross over, p_1 must be on g_1's upper switch-wire output. Accordingly, even the states in which the observation at g_4 was negative have successors that exhibit no interference, as shown by the fact that p_1 emerges from g_3's lower switch-wire output with the expected nonzero frequency following a negative observation at g_4.

Mauritius Renninger (see deBroglie 1964) cites negative observation to demonstrate the incorrectness of one naive account of eliminating interference via observation—the account that attributes this elimination to the inevitable disturbance of an observed entity by the observer. But even though a negative observation can cause no such disturbance (since there is no interaction at all), the interference disappears all the same. Looking at the situation from configuration space, this is just as we would expect: the fact that p_2 encounters p_1 in one of two superposed states makes those two states differ along their p_2-position dimension, moving them out of interference range of one another, thus circumventing interference in both states.

At this point, the quantum interference–observation duality becomes a comprehensible—indeed, deducible—property of the quantish universe. The quantish physical laws say that the configuration-space destination of a classical state's weight is determined only by that state; other superposed states are irrelevant. Therefore, states that are separated from one another along some particle-position dimension in configuration space can interfere with one another only by reconverging to the same point in configuration space (as happens, for example, in fig. 4.12).

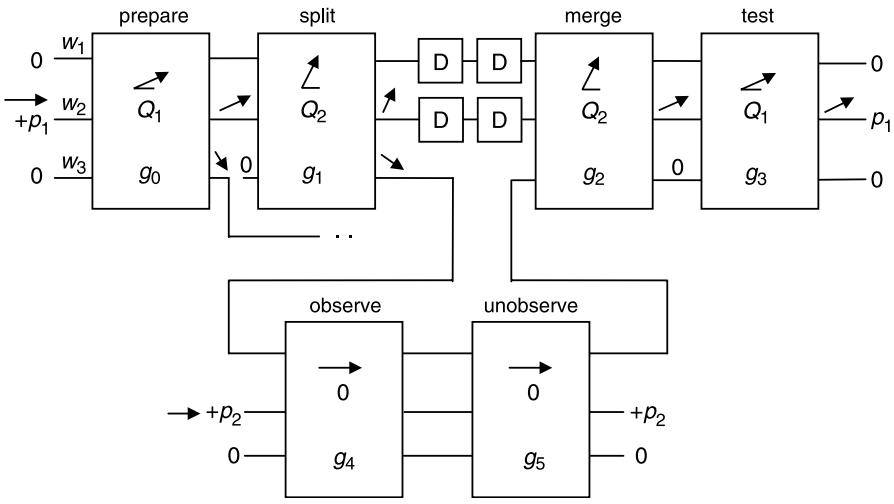


Figure 4.15
Particle p_2 must cross over 0 or 2 times at g_4 and g_5. Either way, it emerges from g_5 at the upper switch wire, erasing the observation made at g_4 and thus reestablishing interference at g_2, as manifested by the definite outcome there and at g_3 (compare the previous figure).

Any observation that distinguishes the superposed states must (as in fig. 4.14) create a corresponding separation along a distinct dimension in configuration space (and any additional such observations, or any observations of the observations, compound the separation along still other dimensions). Then, reversing the original separation creates no interference since there is still separation in one or more *other* dimensions. (But if those other separations are also reversed, interference is indeed reestablished, as in fig. 4.15.) Thus, given the laws of quantish physics, there is a necessary tradeoff between an interfering superposition and any observation that distinguishes among the superposed states.

Thus, the quantish universe—like the real quantum universe—behaves classically to just the extent that we try to catch it in the act of behaving otherwise. The quantish-physics formalism shows how such behavior can be exhibited by deterministic mechanical laws that support only local interactions and that have no peculiarity with respect to there being a definite, objective, observer-independent (quantum) state of the universe. The following section shows that the quantish formalism also supports an analogue of the crucial EPR experiment.

4.3.4 Disproving Hidden-Variable Theories: The EPR Experiment

As a final example, this section presents the quantish parallel of the Einstein–Podolsky–Rosen (EPR) experiment (Einstein 1935). The experiment explores the possibility of *hidden-variable* explanations of quantum mechanics—explanations that postulate that there is no real superposition of distinct states, but rather a single definite state that is merely unknown. (Readers feeling bogged down in the technicalities can safely skip this section; the essential points have already been made.)

Einstein famously recoiled from the nondeterminism of the Copenhagen interpretation. ''God does not play dice with the universe'' was Einstein's metaphorical comment. The EPR experiment was conceived as a way to show whether quantum phenomena involve definite but hidden variables (as Einstein hoped), or else genuine superpositions of coexisting states. The experiment sets up a pair of mutually distant particles whose states are in a correlated superposition (that is, one element of the superposition has both particles in some state A; the other element has both particles in state B).

The experiment was carried out many years after it was first proposed; superpositions won. Hidden-variable theories have been proven false (unless the universe somehow manages to propagate information faster than light and in just the right way to fake a superposition, which is desperately untenable).

Here is the quantish-physics analogue of the EPR experiment. In figure 4.16, p_3 compares p_1's position to p_2's. If both particles have emerged from the upper switch-wire outputs of the splitting gates g_1 and g_2, p_3 encounters both particles and crosses over at both g_3 and g_4. If p_1 and p_2 both emerge from the splitting gates' lower outputs, p_3 encounters neither and passes straight across g_3 and g_4. In either case, p_3 emerges from g_4's upper switch-wire output. But if p_1 and p_2 do not emerge from the corresponding outputs of their respective gates, p_3 crosses over just once (at either g_3 or g_4), and thus emerges from g_4's lower output.

Let us say that p_1 and p_2 are *coupled* in those states in which p_3 has emerged from g_4's upper switch wire, indicating that p_1 and p_2 both emerged from upper switch-wires, or both from lower. The following discussion concerns only the states in which p_1 and p_2 are coupled. (The coupled states do not interfere with the other, uncoupled states, or vice versa, due to their separation in configuration space along the p_3-position dimension.)

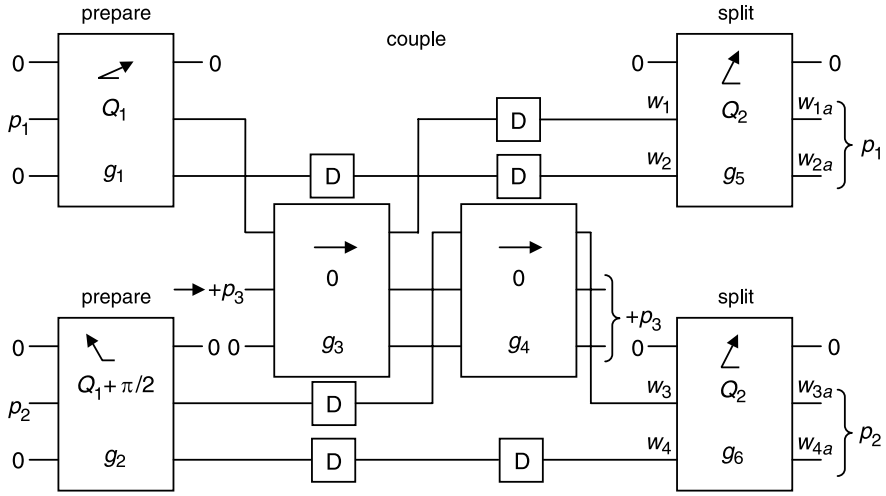


Figure 4.16
The observations by p_3 at g_3 and g_4 couple p_1 with p_2.

Among the coupled states, neither p_1 nor p_2 has a definite position. Rather, each is in a superposition of positions. But that superposition is definite regarding the correspondence of the two particles' positions: each is on an upper wire if and only if the other is too. That is, we have a superposition between two sets of states: one set where both particles are on an upper wire (with various permutations of the particles' signs), the other set where both particles are on a lower wire.

At gates g_5 and g_6, p_1 and p_2 (respectively) encounter measurement angle Q_2 (oblique to Q_1). The outcome is remarkable: regardless of the value of the shared measurement angle Q_2, p_1 and p_2 remain coupled, both emerging from the upper wires or both from the lower wires of their respective gates g_5 and g_6.

The continued coupling can be explained in detail by applying the quantish-physics rules at each of the gates g_1 through g_6, but the gist of the phenomenon is this:

- Following the coupling at g_3 and g_4, the quantum weights at the states in which both coupled particles are at the upper wires turn out to be identical to the quantum weights at the states in which both particles are at the lower wires (fig. 4.17a). In states where p_1 and p_2 pass straight across at g_1

and g_2, the weight is parallel to the original weight rotated by Q_1 at g_1 and further rotated by $Q_1 + \pi/2$ at g_2. In states where p_1 and p_2 cross over instead, the weight is parallel to the original weight rotated by $Q_1 + \pi/2$ at g_1 and further by Q_1 at g_2. Thus, in either case, the two gates together lead to a new weight that is parallel to the original weight rotated by $2Q_1 + \pi/2$. (The initial weight, prior to p_1 and p_2's preparation at g_1 and g_2, is presumed here to have angle 0; any other initial orientation would just rotate all the successor weights by the same amount.)

▪ Due to that symmetrical arrangement of the quantum weights, it then turns out that measuring p_1 at g_5 with respect to Q_2 causes the weights to split and reconverge to the same set of weights that would result from instead measuring p_2 at g_6 with respect to Q_2. Measuring p_1 at g_5 (fig. 4.17b) splits state s_1's weight between s_3 and s_5 along the p_1-position dimension, and splits s_2's weight between s_4 and s_6, also along the p_1-position dimension. Then (fig. 4.17c), measuring p_2 at g_6 splits s_3 and s_4 along the p_2-position dimension (merging at s_7 and s_8), and splits s_5 and s_6 along the p_2-position dimension (merging at s_9 and s_{10}). The weights at s_8 and s_9 cancel to zero, and the weights at s_7 and s_{10} reconstruct the before-measurement configuration (but with the particles shifted to the output wires of g_5 and g_6).

Consequently, when g_6 measures p_2, it is as though g_6 had already measured p_2 (even though instead it was g_5 that measured p_1); measuring p_1 at g_5, and then measuring p_2 at g_6, is like measuring p_2 at g_6 twice in succession (or, equivalently, like measuring p_1 at g_5 twice in succession).

As with any two quantish gates, the state-splitting achieved by g_5 and g_6 simultaneously is the same as if the splitting occurred first at g_5 and then at g_6, or vice versa. We just imagined that the first split occurs at g_5. If instead we construe the two simultaneous measurements to be performed in the opposite order, then the p_2-position split precedes the p_1-position split. But then figure 4.17b still looks the same as before (except that the p_2 positions have changed instead of the p_1 positions); and the final configuration in figure 4.17c remains identical to what is shown.

As we have already seen (fig. 4.12), measuring a particle twice in succession (using the same measurement angle and without diverting the particle from either switch wire) reconverges the split weights and reconstructs the original definite state. And as just remarked, because of the coupling here,

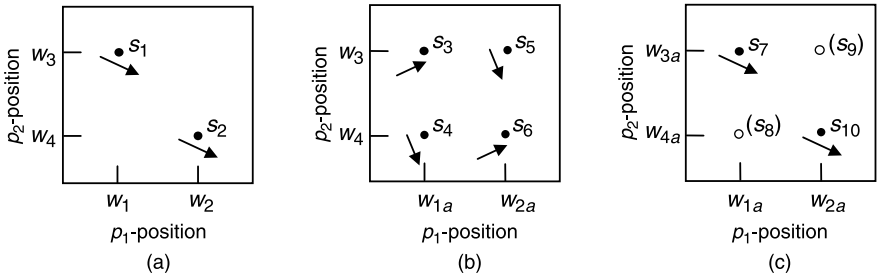


Figure 4.17

Configuration-space view: Each superposed state s_i shown here is actually a disjunction of four more-specific states, one for each permutation of the two particles' signs. Each s_i is shown along with the quantum weight there—actually the sum of the four constituent states' weights. As usual, the weights' orientations are depicted in the complex plane.

measuring both particles with respect to the same angle at g_5 and g_6 is like measuring the same particle twice with respect to that angle. Accordingly, the weights reconstruct the original definite state with respect to both particles, and p_1 emerges from the upper switch wire of g_5 if and only if p_2 emerges from the upper switch wire of g_6. (In contrast, of course, if gates g_3 and g_4, and the associated delay gates, had been omitted from the circuit, then there would be also be nonzero-weighted states in which p_1 emerges at g_5's upper switch wire and p_2 emerges at g_6's lower switch wire, and vice versa.)

That the particles thus remain correlated after being measured by g_5 and g_6 can be demonstrated from within the quantish universe by observing both particles' emergence from those gates over a large number of trials, and verifying that both always emerge from the upper switch wires, or else both from the lower switch wires. (After the observations, the particles proceed independently, no longer coupled with respect to any subsequent measurements.) Using various values for Q_2 on different trials shows that the phenomenon holds for all measurement angles.

But the *indefiniteness* of the particles' states (prior to measurement) is harder to show from within the quantish world. Proponents of a classical worldview—a view that denies the reality of multiple superposed states of the universe—would seek to explain the demonstrated correspondence of the measurement results at g_5 and g_6 by postulating that from the outset

of the experiment, prior to the comparison performed by p_3, p_1 and p_2 each already had a definite (albeit unknown) state for every measurement angle (thus violating the quantish analogue of Heisenberg's uncertainty principle, as discussed in sec. 4.3.1). This classical view proposes that on each trial, both particles start with the *same* definite (although hidden) state, thus explaining the later-observed correspondence for any measurement angle. This is the view that Einstein hoped (in vain) would be supported by the outcome of the EPR experiment.

From our privileged vantage point, looking in from outside the quantish universe, we know that this so-called hidden-variable account is false. We see that configuration space provides a superposition of outcomes at both gates, not a single, definite outcome at each. But can the hidden-variable interpretation of the EPR experiment be disproved from within the quantish universe? A subtle theorem due to Bell (1964) facilitates such a proof.

Gate g_5 measures p_1 with respect to Q_2. What is measured is p_1's inclination to pass straight across or to cross over (given angle Q_2)—a binary attribute of p_1. Suppose we modify the experiment to substitute a distinct angle Q_3 for Q_2 at g_6. Thus, g_6 now measures p_2 with respect to Q_3. Let us define the *discrepancy rate* between the measurements at g_5 and g_6 as the (apparent) probability that, after passing through those gates, p_1 and p_2 will not both be on upper wires or both on lower wires (where apparent probability is as seen from within the quantum universe, by keeping a cumulative record of the outcomes of many trials of the experiment). As noted above, g_5's measurement of p_1 is effectively as though g_6 had measured p_2 with respect to the same angle Q_2. Therefore, p_2's measurement with respect to Q_3 occurs just as though p_2 itself (rather than p_1) had already been measured with respect to Q_2. As in figure 4.11, that sequence of measurements has a discrepancy rate of $\sin^2(Q_3–Q_2)$—which, of course, is zero if $Q_2 = Q_3$, as in the initial discussion above.

Next, consider whether the observed discrepancy rates for various values of Q_2 and Q_3 are explicable by postulating that the particles have prior definite states for the Q_2 and Q_3 measurements. *Bell's theorem* states that

- if each pair of coupled particles already has a single definite state for each of three arbitrary measurement angles Q_a, Q_b, and Q_c, and those definite values are assigned independently of which measurements will in fact be performed;

- and if we perform measurements on many pairs of coupled particles;
- then the discrepancy between the Q_a and Q_c measurements (that is, the discrepancy rate among trials in which one coupled particle is measured with respect to Q_a, and the other with respect to Q_c) cannot exceed the discrepancy between the Q_a and Q_b measurements plus the discrepancy between the Q_b and Q_c measurements. This comparison is *Bell's inequality*.

Bell's inequality follows simply from the fact that any particle with a different state with respect to Q_a than with respect to Q_c must also have a difference between its Q_a and Q_b states or between its Q_b and Q_c states—since its Q_b state cannot match both its Q_a state and its Q_c state if its Q_a and Q_c states differ.

Let us take Q_a to be 0, Q_b to be $\pi/8$, and Q_c to be $\pi/4$. If we perform a series of experiments in the quantish universe, using the setup of figure 4.16, variously choosing the values of Q_2 and Q_3 from Q_a, Q_b, and Q_c, we will find that the discrepancy between Q_a and Q_c is $\sin^2(\pi/4) = 0.5$, and the discrepancy between Q_a and Q_b, and also between Q_b and Q_c, is $\sin^2 \pi/8$, which is about 0.146. This result clearly violates Bell's inequality.

Therefore, the observed correlation between paired particles' measurements with respect to angles 0, $\pi/8$, and $\pi/4$ cannot be explained by saying that on each trial, for each possible measurement angle, the two particles already shared a single definite state (thereby accounting for the exceptionless match whenever the two particles are measured with respect to the same angle). By Bell's theorem, that interpretation is impossible.

If one were to deny the reality of multiple superposed states of the universe, the only remaining way to account for the observed correlations among the coupled particles' measurements with respect to the three angles would be to postulate that the outcome of measuring one particle is—by some unknown, unexplained mechanism—then communicated to the other coupled particle, in such a way as to force the other particle into the same state with respect to whatever measurement angle was used for the first particle. In any experiment that creates coupled states, this communication would always somehow occur, always somehow conspiring to simulate the requisite quantum superposition.

In fact, of course, no such communication is involved, nor could there be, given quantish physical laws, when there is no circuitry between the two measuring gates to transmit the outcome from one gate to the other.

The quantish-physics model instead accounts for the correlation by saying that there is a superposition of appropriately weighted entire classical states of the universe. Interference among the superposed states creates correlations that would be impossible, by Bell's theorem, if there were only one such state, in the absence of an avenue of communication.

The foregoing is adapted from the proposed EPR experiment, later carried out in modified form by several investigators (e.g., Aspect, Dalibard, and Roger [1982]). The measurements of the two particles are performed far enough from one another in space, yet closely enough in time, that anyone who rejects the reality of multiple superposed states of the universe is thereby forced to postulate an unexplained, faster-than-light interaction.

The supposed faster-than-light interaction would not only be contrary to what is ever directly demonstrated, but furthermore, as just noted, it would always take place in such a way as to simulate a correlated quantum superposition whenever particles are in coupled states. Yet somehow, there is no way to harness this putative interaction to communicate a message from one coupled particle to the other—the two measurement results always match each other, but there is no way for either to pin down which result the other will share (because in reality there is a superposition of different shared results). This curiosity is just what we would expect were there not, in fact, an interaction, but rather a manifestation of a preestablished correspondence—though not a correspondence between definite states of the two particles (as hidden-variable theories propose), but rather a correspondence between the two particles that exists within each of two superposed states.

Thus, the putative faster-than-light interaction is less plausible than the reality of the superposition itself—especially since the observation–interference duality of the double-slit experiment (as recapitulated in the quantish world in sec. 4.3.3) is already paradoxical with respect to hidden-variable accounts and already requires appeal to quantum superposition to resolve the paradox. In the double-slit experiment, hidden-variable accounts are paradoxical in that they require a particle to quantitatively mimic being subject to quantum interference with superposed versions of itself, but only to the extent that no observation distinguishes the elements of the superposition. This mimicry would require a particle's mysterious, systematic response to an open-ended variety of possible remote events—events such as an observation of the path the particle did *not* take. The

EPR experiment just adds the twist that the mysterious systematic response would also have to propagate faster than light, making it even less tenable.

4.4 Many Worlds or Quantum Collapse?

The multiplicity of outcomes in Everett's interpretation of quantum mechanics stems from what we might call the *contagion* of a particle's superposition when its state is observed by another particle[11]—the other not only enters a superposition of states, but it enters a correlated superposition. The resulting quantum state therefore cannot be expressed simply as the cross product of two independent superpositions, but must instead designate superpositions of configurations of both particles. And as observations of observations cascade, arbitrarily many particles may join the correlated superposition, effectively splitting the universe into separate versions, at least as far as the participating particles are concerned.

An elegant formalism for this process, explored here as an analogue of Everett's formulation, is to represent the quantum universe in terms of weights on total classical states of the universe. These weights flow deterministically through configuration space, and a so-called split occasioned by an observation is merely the further separation, along additional configuration-space dimensions, of two quantum weights attached to two already-distinct classical states, already separated along one or more configuration-space dimensions.

The Copenhagen interpretation is almost identical to Everett's. In particular, the contagion of superposition when particles interact is also present in the Copenhagen formalism, which is, indeed, identical to Everett's formalism. The contagion of superposition is what explains the quantum hide-and-seek game, providing a correlation between the observer and the observed, and providing an inherent complementarity between that correlation and quantum interference.

But the Copenhagen interpretation, unlike Everett's, diverges from the formalism by positing an extra event—the collapse of the superposition onto just one of its superposed states—that contradicts the formalism. According to the Copenhagen interpretation, this collapse occurs at some

11. Physicists use the term *entanglement*.

unspecified point along the cascade of microscopic observations, so that at least by the time the observations culminate in a conscious observation by a human being (or perhaps even by the time they culminate in a macroscopic observation by, say, a laboratory instrument), the superposition has collapsed. Additionally, the Copenhagen interpretation tries to distance itself from the formalism by insisting that the superposed states described by the formalism are not yet real, just potential ones. Only one such state may eventually become real, when an observation collapses the superposition onto that state. But in view of the superposed states' acknowledged influence on real events, withholding the label ''real'' from the superposed states is an empty gesture.

The chief motivation for postulating the collapse is straightforward. Following a quantum experiment, the formalism predicts a continuing superposition of states. But the experimenter clearly observes only one state from that superposition. Therefore, the superposition has apparently collapsed into a unique state, contrary to the formalism. Everett's central contribution was to demonstrate that the formalism already accounts for the seemingly unique outcome, for although the formalism describes a continuing superposition of states, it also describes a corresponding superposition of mutually isolated observers, each of whom will indeed see only one outcome.

A common paraphrase of Everett's formulation is that an act of observation splits the entire universe into distinct branches, all of which persist; many authors call this the *many-worlds* interpretation of quantum mechanics (DeWitt and Graham 1973). Stated thus, Everett's interpretation of observation sounds like a desperate, ad hoc complication of Schrödinger's formalism, motivated only by wanting to avoid the Copenhagen interpretation's competing complication, the nondeterministic collapse. The ensuing battle pits the determinist intuition against the unitary-universe intuition—a contest in which the Copenhagen interpretation easily holds its own.

But contrary to what the common paraphrase may suggest, Everett's multiplicity of worlds does not arise all at once as an enormous global response to every microscopic observation. Rather, exactly as in the Copenhagen interpretation, it proceeds one particle at a time, as each next observing particle interacts with and becomes correlated with an initial

superposition. That is, with each such interaction, the superposed states move apart along yet another configuration-space dimension, a dimension corresponding to the newly participating particle.

In an important sense, by Everett's interpretation, the entire universe may as well have already split up, since any part of it that ever observes the already-superposed part will, at that point, participate in the split. But the actual splitting is incremental, not global, even if it *eventually* involves all the particles in the universe. And the split occurs under either quantum interpretation—until, according to the Copenhagen interpretation, the superposition supposedly collapses. Accordingly, I avoid the popular many-worlds label for Everett's view, instead using his original term: the relative-state interpretation.

Thus, contrary to how the debate is often framed, the difference between the Copenhagen and Everett interpretations is not a dispute between a single branch and multiple branches. The multiplicity of branches, in the sense of the contagion of superposition when particles interact, is a property of the formalism that both interpretations share.

And let's be clear: this shared property is a spectacularly, unprecedentedly bizarre contradiction of our everyday, commonsense perception of the world. A theory so bizarre deserves to be met with extreme skepticism, until and unless proof of the theory becomes overwhelming. But quantum mechanics—including the contagion of quantum superposition—*is* overwhelmingly supported by more than a century of science and technology. It is not only demonstrated reliably and repeatedly in the laboratory, but it forms the basis for various electronic devices we use every day.

Thus, a truly bizarre multiplicity of superposed quantum branches has long been incontrovertibly established, and is common to both major competing interpretations of quantum mechanics. The only difference between the interpretations is whether or not to postulate an extra kind of event, the eventual collapse of the superposition.

In view of Everett's explanation, the seemingly unique outcome of a quantum observation does not provide evidence for a collapse of a quantum superposition. Even without the collapse, Everett explains why you—that is, each version of you—will see only one outcome. Thus, with regard to the seemingly unique outcome, the collapse becomes a superfluous

hypothesis, just like the gravity-deflection-of-photons explanation in section 1.2.3's mirror-asymmetry paradox. But (in contrast with the gravity-deflection notion) there exists no other evidence to show that there has ever been such a collapse, to any extent, under any circumstances. Postulating the collapse thus becomes a gratuitous complication and contradiction of the massively confirmed formalism. Moreover, the collapse renders quantum theory incomplete and ambiguous:

- The theory becomes incomplete because it cannot describe a quantum state of some portion of the universe, except relative to some other portion that embodies an observer. The theory cannot, in principle, describe the quantum state of the universe as a whole, and give laws for the evolution of that state. Everett's relative-state formulation can and does.
- The theory becomes ambiguous regarding what sort of physical interaction constitutes a superposition-collapsing observation. Yet the theory makes different predictions depending on whether such an observation has occurred: in particular, if the observation is later ''reversed,'' reconverging the superposed states (as in fig. 4.15 in sec. 4.3.3), interference occurs if the superposition is intact, but cannot occur if the superposition had collapsed, leaving nothing to interfere with. (But Copenhagenists only postulate a collapse when reversal is prohibitively unlikely, so that the distinguishing experiment is prohibitively impractical.)

Not only is the postulated collapse unsupported by any evidence, and incompletely and ambiguously specified, but furthermore, the most problematic features of quantum physics—the apparent nonobjectivity of the state of the universe, apparent nonlocality of the effects of a measurement, and apparent nondeterminism—result from postulating the collapse. Since the formal model already accounts for all the facts, what motivates adding the quantum collapse?

Here are some key arguments in opposition to Everett's alternative to the quantum collapse.

The Argument from Noninteraction Following a macroscopic quantum observation—by a person, say, or a laboratory instrument—branches of the universe may diverge irreversibly (thermodynamically irreversibly: it is intractably unlikely that every microscopic manifestation of the observation will be undone, leaving no trace of the observation). If there is a

branch of the universe with which we can never interact, we may ask what justifies calling it real. We would not, after all, acknowledge the reality of some arbitrary, fanciful entity that putatively sits in our midst but is forever impossible to detect.

But, as argued above, other branches of the universe do interact with ours—quantum interference *is* that interaction. True, when a practically irreversible quantum observation occurs, a previously interacting branch of the universe moves, as a practical matter, forever ''beyond the range'' of further nonnegligible interaction (physicists call this separation *decoherence*). But to say that the branch thereby ceases to exist makes no more sense than to deny the continued existence of a reflected photon once the photon has become so distant from us that as a practical matter, we cannot hope ever to detect it again. Such a stance is blatant solipsism. Demonstrably real things do not cease to exist merely because they move too far away—in physical space, or in configuration space—for us to be likely to see them anymore. Denying a receding photon's continued existence would (among other problems) contradict the conservation of matter and energy. Denying a receding universe-branch's existence wreaks similar havoc; as already noted, that denial ends up contradicting the determinism and locality of physical laws, and even the observer-independent objective existence of the universe's contents.

Moreover, as argued in a groundbreaking paper by David Deutsch (1986), radically divergent branches of the universe can, in fact, reconverge. We can construct a *quantum computer* in which arbitrarily complex observations are erasable in the manner of figure 4.15. Recently, an early prototype of a quantum computer was built (Vandersypen 2001). It can find the factors of an integer by exploring different pieces of the solution simultaneously in different branches of configuration space, and then reconverging the branches to assemble the solution via those branches' mutual interference. In each branch, particles at various quantum-logic gates are in a superposed quantum state specific to that branch's piece of the solution. The particles' states are observed by particles at other gates that compute some function of the observed states. Later, the observations are undone to reconverge the branches.

If the apparent quantum collapse entailed the disappearance of all but one superposed state when such observations occur, there would be nothing left to reconverge with when the observations are later undone. Thus,

to account for reconvergence, the Copenhagen interpretation has to deny that a later-undone observation was in fact an observation, even though it would have been were it not going to be undone. Only because the Copenhagen interpretation is unconstrained by any definite principle regarding what constitutes an observation can it make this denial without contradiction.

The Argument from Immensity The configuration-space universe is, to put it mildly, very big. It seems somehow counterintuitively wasteful for there to be exponentially many versions of the world when a single one would do just as well. I must confess to feeling the tug of this intuition myself—though curiously, my intuition balks more at the size of a perhaps-finite configuration-space universe than at a perhaps-infinite classical universe.

But this intuition is surely an expression of naive anthropocentrism: only if the "purpose" of the universe were to support middle-sized objects such as us would it be surprising to find that things are far more complicated than that purpose might logically require. It would be no less plausible to deny the reality of atoms on the grounds that unimaginably many of *them* are required to explain the simple, everyday phenomena we perceive.

Of course, parsimony is a legitimate criterion for judging theories. But what counts is parsimony of explanatory concepts, not (or not as much) parsimony of the objects being explained. If a compact set of laws accurately describes a universe of vast intricacy, so much the better for those laws. We should not posit an extra kind of event in the universe—especially an event for which there is no evidence, an event that has no precise description and that would violate locality, determinism, and the objective nature of reality—merely because the consequent universe would be much smaller.

The Preferred-Basis Argument Recall in section 4.3.1 that a superposition with regard to one measurement angle can equivalently be viewed as a different superposition with regard to another angle. Each possible measurement angle corresponds to a distinct so-called *basis*, distinguishing between a measurement-parallel and measurement-perpendicular component. Sometimes, a given measurement angle corresponds to a perspicuous partitioning of configuration space, as for example when a particle has a

definite value with respect to a particular angle. But then there is necessarily a multiplicity of values with respect to other (oblique) measurement angles.

Similarly, in real-world quantum physics, a superposition that neatly divides configuration space into, say, two regions—one in which we observe a given quantum-measurement outcome, and one in which we observe a different outcome—could alternatively be described, with respect to a different basis, as a messier superposition of two states, each of which consists partly of one outcome and its observations, and partly of the other outcome and its observations.

If universe-splitting were a decisive event in which two or more branches of the universe are explicitly designated as separate from one another, then we would need to explain why the split occurs with respect to the basis that yields a convenient dichotomy. The quantum formalism itself does not pick out any such preferred basis. Thus, some skeptics (e.g., Kent 1990) have argued that the formalism is incomplete—that whether or not there is an actual collapse, there is at any rate something going on in the apparent collapse that Everett's interpretation does not account for.

But again, Everett's interpretation of a quantum measurement neither proposes nor requires any extra event of universe-splitting. Instead, Everett merely notes that one correct description of the quantum-amplitude distribution (among many such descriptions, for different choices of a basis) is that we have a pair of universe branches, one for each outcome, with observers in each branch seeing that branch's outcome. That is, the quantum formalism is consistent with having such a pair of branches.

True, the formalism is equally consistent with other, less perspicuous pairs of branches. Nothing, though, requires us to anoint the perspicuous basis the sole ''real'' one. True, other basis choices may not coherently describe conscious observers at all. Since there obviously are conscious observers, doesn't that require the ''real'' basis to support them? No, it just requires *a* basis to support them. That's all it takes for them to be physically real, for them to consist of real physical events, if consciousness is just a particular kind of physically ordinary mechanical process, as discussed in chapter 2.[12]

12. Wallace (2003a) argues in more detail for the superfluousness of a preferred ''real'' basis.

The Argument from Unitary Consciousness Some people can accept the immensity of the quantum superposition of universe branches, but reject the idea that they themselves might split into multiple versions. The paradox seems to be: when I look at an apparatus that measures the outcome of a quantum event, if the universe then splits into two branches, with two versions of myself, which version of myself do "I" become? Each version next sees a different state of the apparatus. But which is "my" next experience? Clearly not both of them (I never see both outcomes together); but if only one, then is that not the only real outcome?

This is indeed paradoxical if one conceives of time flowing forward, with one's consciousness flowing along with it from temporal version to temporal version of oneself. Alternatively, though, we can acknowledge that there is no such flow. Everything is just sitting statically in spacetime; each next temporal version of oneself includes memories of previous versions, and thus experiences the illusion of a unitary consciousness having been passed to this version from the previous versions. This is precisely the view already argued for in chapters 2 and 3. From this point of view, there is no paradox if multiple subsequent versions, instead of just one, happen to share the illusion of a unitary, forward-flowing consciousness.[13]

Some authors[14] have objected that Everett's formulation does not explain why you are conscious of only one outcome of a binary quantum measurement when you observe the experiment's outcome. After all, even *assuming* a preferred basis that neatly divides configuration space into a pair of outcomes and observers, the quantum equations still dictate a superposition of observers, one for each outcome. Why, then, is there not a still-unitary consciousness that embraces both observations, rather than a bifurcation into two separate consciousnesses, each seeing only one outcome?

From a dualist standpoint, if consciousness were some ghostlike entity somehow generated by or attached to our physical brains, we would indeed lack an answer to that question. Why, after all, could the consciousness not

13. Philosophers grapple with similar paradoxes concerning a transporter beam—a hypothetical science-fiction device for disassembling a person's atoms and assembling an exact copy at a remote location. If *two* such copies are assembled (or if the original is left intact), the same questions arise about consciousness as are posed here by split universes (see Nozick 1981, chap. 1).

14. See, e.g., Penrose 1989, p. 296.

just attach to the whole superposition? But if consciousness is a physically ordinary computational process—perhaps involving something like a Cartesian Camcorder's smart recording and playback of some thoughts and perceptions, as in chapter 2—then Everett's formulation straightforwardly answers the question. The quantum laws give us two separate, noninteracting collections of particles following a quantum observation. The two versions of the observer operate entirely independently of one another, with neither version able to see or record or play back what the other sees.

We do not need to know precisely how consciousness works (and indeed we do not yet know that) in order to know that *if* consciousness is a physically ordinary mechanical, computational process, then Everett's interpretation explains why there are two separate versions of the observer, each observing only one outcome. Asking why the superposition of observers does not constitute a single overarching consciousness of both observations is like asking why a superposition of two photographs of the outcomes does not constitute a single double-exposure showing both outcomes. We need not know the chemical details of how film photography works to know that the quantum-superposed photographs are as separate from one another as are any two physically distinct photographs that coexist in the *same* quantum branch. For example, it is possible to shred one of the superposed photographs while the other remains intact, something that cannot occur with the two images of a double exposure—and similarly for the two superposed versions of a conscious observer.[15]

The unitary-consciousness objection to universe splitting thus has nothing to do with physics per se, and everything to do with one's conception of consciousness. Here again, we see the mutual ramifications of far-flung aspects of the question of mechanical reality. Precisely at the point where universe branches diverge for a quantum observation made by a conscious

15. As mentioned above, it is only a particular choice of a configuration-space basis that divides the world into two branches, each of which contains one version of a conscious observer. Other bases instead yield two mishmashes not only of versions of an observer, but also of versions of a photograph or any other physical object. If reality consists ultimately of quantum amplitude flowing through configuration space, then a person, photograph, or other object has physical reality insofar as there is *some* configuration-space basis in terms of which the object in question is implemented.

observer, some nonmechanical concepts of consciousness come into direct conflict with physical principles.

For physicists such as von Neumann, it is the physical principles that yield. To accommodate his unwillingness to accept a splitting of mind, von Neumann denied the entire splitting of the physical world, postulating instead a quantum collapse at just the point where conscious observation occurs (just as, in the mirror paradox of section 1.2.3, one might be tempted to postulate underlying physical processes that distinguish the horizontal from the vertical, to account for those dimensions' seemingly asymmetric treatment when we look in a mirror). Other physicists, skeptical of that intrusion of consciousness into physics, assume that the collapse can happen earlier, as a result of (some) inanimate observations. But they still presume that the collapse occurs *at least* by the time of conscious observation.

Thus, contrary to a popular perception, the apparent role of consciousness in the putative quantum collapse is not suggested by any physical evidence (nor does any physical evidence even suggest the collapse itself). Rather, some physicists' prior conception of consciousness (at least with regard to the prospect of its splitting) distorts their interpretation of the physical evidence. If the distortion goes unnoticed—if the supposedly unitary, unsplittable nature of consciousness just seems too obvious to question—then the presumption about consciousness gets smuggled into physics undetected, and the resulting version of physics then magically seems to support the notion of a special physical role for consciousness.

4.5 Summary

Everett's interpretation of quantum mechanics merely takes seriously the formalism that accords with all our experimental experience; Everett just regards the formalism as describing physical reality. In this interpretation, there is no such event as the collapse of the superposition. Instead, superposed states all persist, but may move out of range of interacting with one another. Conscious observers are not physically special, and are subject to the same quantum superposition as any other physical objects. The resulting physical system is straightforwardly objective and mechanical, even local (i.e., no instant action-at-a-distance) and deterministic (though not in the way that hidden-variable theories propose).

There is no good reason to postulate the quantum collapse. But there are several seemingly good reasons to do so, at least one of which is deeply rooted in traditional, prescientific beliefs about the nature of consciousness. The force of these flawed reasons may explain why many physicists still postulate the collapse.

Quantish physics—although not identical to actual quantum physics—shows, by example, how it could be that local, deterministic laws produce a quantumlike interference–observation duality. Everett's formulation does the same, for a more complicated example: the real world.

Everett's interpretation of quantum mechanics accounts for actual quantum phenomena in terms of an elegant formalism—one so basic to the phenomena that even the Copenhagen interpretation invokes that formalism, together with a gratuitous complication, the quantum collapse. Everett's model *explains* the quantum interference–observation duality—the duality can be deduced from the model—instead of requiring an ad hoc, imprecise distinction between observation interactions and all other physical interactions.

The quantish-physics model is much simpler than, but deeply similar to, Everett's formulation. Quantish physics faithfully exhibits not only the fundamental interference–observation duality, but also (with respect to definite states) Heisenberg's impossibility of eliminating a superposition without thereby introducing some other superposition. Quantish physics even supports an analogue of the EPR experiment, which shows that quantum superposition cannot be explained by hidden variables—that is, by definite, unknown but preestablished states each of which provides for a unique measurement outcome.

By substituting trivial Fredkin-gate mechanics for real-world wave mechanics, quantish physics allows us to devote full attention to what is special and perplexing about quantum uncertainty. The quantish model captures the fundamental issues that the interpretational debate appeals to, and captures them in a precise formalism that can be understood without the training required of physicists.

And this accessibility matters, for more than the usual reasons of scientific literacy. The difference between a mechanistic and nonmechanistic universe is as profound a philosophical matter as we have ever grappled with. To the extent that that dispute focuses on quantum physics, a simplified model such as quantish physics may provide a common ground on

which laypersons, philosophers who are not physicists, and physicists who are not philosophers can communicate with precision.

It is important to know what kind of world we inhabit—a world described and explained by mechanical principles, not by supernatural spirits—because it is important to understand how to find the truth: by rational inquiry, not by appeal to mystical authority. The Copenhagen interpretation of quantum mechanics is incoherent and thereby makes reality itself seem incoherent. Nothing could be more profoundly mistaken.

5 Deterministic Choice, Part 1: Inalterability Does Not Imply Futility

The last two chapters focused on aspects of the mechanical-universe paradigm that seem paradoxical from the standpoint of the outside world itself: the apparent forward flow of time, though physical laws prescribe no such flow; and the apparent quantum-mechanical hide-and-seek game whereby multiple superposed states sometimes seem to coexist (as is shown statistically by their mutual interference over many trials), although only one of them (at random) is ever seen when we look. In both cases, resolving the paradox required careful consideration of the status we observers have as mechanical entities that are part of the physical system under examination.

In this chapter and the next two, I turn the spotlight away from the rest of the universe and back squarely in our own direction. Whereas the previous two chapters examined our status as observers, I now turn to our status as *actors*, that is, as agents who initiate actions. I take up two threads left dangling from chapter 2's discussion of consciousness:

- This chapter investigates the *paradox of choice*: how can it be that we make choices for ourselves if we are just machines, completely constrained by the physics of our constituent parts? What difference could choice even make, if (as discussed in chaps. 3 and 4) the universe is deterministic, with everything (including the entire future) already just sitting statically in spacetime?
- Chapter 6 sharpens and extends the analysis of choice by looking at Newcomb's Problem (introduced briefly in sec. 5.1 below).
- Chapter 7 investigates the question of the foundation of value and ethics: if we are just collections of atoms, then how can it matter what happens to us? How can there be a right or wrong about how we treat one another?

The question of the nature of choice and the question of the foundation of value and ethics are perhaps-surprisingly interlinked via a scenario

known as the *Prisoner's Dilemma*. In that scenario, each of two prisoners must separately choose either to *defect* by implicating the other, or to *cooperate* with the other by staying silent. They decide independently, with no communication allowed between them. They both know the rules, which are as follows:

- If it turns out that one defects and the other cooperates, then the defector goes free and the cooperator spends life in prison.
- If both cooperate, they both spend five years in prison.
- If both defect, they both spend ten years in prison.

Imagine the situation from the point of view of either one of the prisoners. If the other prisoner defects against you, your sentence is milder if you defect (10 years) than if you cooperate (life). Alternatively, if the other prisoner cooperates with you, you still do better if you defect (you go free) than if you cooperate (5 years). So given the other prisoner's choice—regardless of which choice it is—you do better if you defect than if you cooperate. Defection, then, is seemingly the more beneficial choice for you.

On the other hand, the symmetry between your situation and the other prisoner's assures that whichever choice is more beneficial to you for you to make, the corresponding choice is also more beneficial to the other for the other to make. If defection is your best choice, it's also the other prisoner's best choice. Yet you both do better if you both cooperate (five years) than if you both defect (ten years). Therein lies a paradox. Suppose you are each rational enough to recognize the better choice for yourself and to act accordingly. How, then, can defection be the rational choice for each of you when you would both do better if cooperation were the rational choice? How can the rational choice be the one that leaves you (both) worse off than if the other choice were the rational one?

The next three chapters explore the compatibility of choice with determinism and mechanism, the relation of that compatibility to the Prisoner's Dilemma, and the ramifications concerning a foundation for ethics.

5.1 The Paradox of Choice without Change

If you have a goal that would be achieved if and only if you took a particular action, then (other things being equal) it makes sense for you to take

that action for the sake of the goal. But if there is something your action cannot alter (cannot make different from what it already is), then it is futile for you to act for the sake of its being one way or another. These two innocuous-sounding intuitions normally coexist peacefully. But the thought experiment known as Newcomb's Problem (Nozick 1969) places them in contradiction of one another, exposing a deep underlying conflict.

In Newcomb's Problem, a mischievous benefactor offers you a transparent box containing $1,000. Awhile ago, the benefactor predicted, very reliably, whether you would choose to accept or refuse the transparent box. (Let's make a science-fiction assumption—elaborated on in sec. 6.1 below—that the benefactor was able to run a very fast, very accurate simulation of you and your surroundings to make the prediction.) There is an adjacent, opaque box that is yours no matter what. In it, your benefactor has already placed $1,000,000 for you if the prediction showed you would refuse the transparent box. If the prediction showed otherwise, the opaque box was left empty. The opaque box has been sealed, and its content cannot subsequently change—whichever choice you make, the box content ($0 or $1M) stays the same as whatever it already was before your choice.

You are accurately and convincingly informed of these circumstances. If and only if you were to take just the opaque box, declining the transparent one, you would expect to find $1,000,000 in the opaque box, so that's apparently the right choice. But taking *both* boxes gets you an extra $1,000 and cannot alter whether there's $1,000,000 in the already-sealed opaque box, so *that's* apparently the right choice.

Now the two prescriptive intuitions conflict: *take an action for the sake of what would be the case if and only if you so acted*, versus *don't bother to take an action for the sake of what your action cannot alter*. This conflict bears on the oft-perceived incompatibility between choice and determinism, and also bears on a similar intuitive conflict about the rationality of cooperative action in Prisoner's Dilemma situations:

- For any goal in a deterministic universe, the inalterable past state of the universe (like the inalterable opaque-box content) already guarantees the achievement or nonachievement of the goal, seemingly rendering any action futile.
- Similarly, suppose you and someone else must decide independently whether to cooperate with one another in a Prisoner's Dilemma situation,

where both of you would be better off if you both cooperate than if you both do not, but only the other's cooperation (not your own) causes any benefit to you. Whichever choice is correct for you must be correct for your symmetrically situated counterpart; but from the standpoint of your self-interest, it is seemingly futile for you to cooperate, since the other's decision whether to cooperate is inalterable by your own decision.

I argue here that (quantum mechanics notwithstanding, regardless of whether chap. 4's deterministic interpretation turns out to be correct) our universe is deterministic and predictable *enough—especially* regarding choice—that the nonfutility of our choices can be vindicated if and only if choice would make sense even given full determinism. Hence, it is important to show that choice and determinism are compatible, regardless of how our universe's actual physics turns out. This reconciliation of choice and determinism requires rebutting the fatalist intuition that inalterability implies futility. The case for cooperative or benevolent behavior (even when your own cooperation cannot alter others' causally independent choices), an important matter in its own right, turns out to benefit from the same rebuttal.

In this chapter and the next two, I argue in favor of choice given determinism, in favor of pursuing the $1,000,000 in Newcomb's Problem by declining the $1,000, and in favor of cooperative action in the Prisoner's Dilemma. The discussion proceeds as follows. In the remainder of this chapter, section 5.2 asks under what conditions it makes sense to take a given action for the sake of a given goal. Section 5.3, following, for example, Dennett (1984) and Minsky (1968, 1986), discusses acting in pursuit of goals despite determinism. Section 5.4 argues that it can make sense to act even for a goal that your action does not cause. Section 5.5 argues that, nonetheless, sensible goal-pursuit requires more than a correlation between action and goal, and section 5.6 suggests what (if not causation) the extra ingredient might be.

Then, in the next chapter, section 6.1 analyzes Newcomb's Problem in light of the preceding discussion. (Some readers may be skeptical that the situation posited in Newcomb's Problem is even logically possible; sec. 6.1 argues at length that it is.) Sections 6.2 and 6.3 consider radical variants of Newcomb's Problem in which *both* boxes are transparent. Chapter 7 then tackles the Prisoner's Dilemma and its ethical implications.

5.2 Means–End Relations

Let us say that there is a *means–end* relation between a contemplated action and a goal just in case the desirability of the goal rationally contributes motivation for taking the action—that is, just in case, other things being equal (i.e., in the absence of conflicting consequences of higher priority), it makes sense to take the action for the sake of the goal's achievement.

By what criteria can we recognize the existence of a particular means–end link? (I use *link* as an informal synonym for *relation*.) Prominent suggestions include (in order of increasing strictness):

- *Evidential* or *correlational* criterion: there is a means–end link from action to goal just in case the goal is more likely to be found to obtain when the action is found to be taken than when the action is found not to be taken (i.e., the action's occurrence is correlated with, and thus gives evidence of, the goal's occurrence).
- *Subjunctive* or *counterfactual* criterion: there is a means–end link from action to goal just in case the goal *would* obtain if the action were taken, but not otherwise (or at least the goal would more likely obtain if the action were taken than if otherwise).
- *Causal* criterion: there is a means–end link from action to goal just in case the action causes (or tends to cause) the goal to obtain.
- *Fatalist* criterion: there is never a means–end link from action to goal; all actions are futile. (No one takes fatalism seriously in practice, but many believe it would indeed follow if the universe were deterministic; hence, they reject determinism.)

The second of these criteria requires a bit of elaboration to explain so-called subjunctive (or *counterfactual*) inference. Inference (or *implication*) involves propositions of the form *If X then Y*; we infer consequent *Y* from antecedent *X*. But logicians distinguish several varieties of inference, including in particular subjunctive inference and *material implication*. Mathematical logic more often uses the latter; it is much simpler to formalize. In the sense of material implication, *If X then Y* just means *It is not the case both that X is true and Y is false*.

In English, we often express subjunctive inference in the subjunctive tense (*If X were true, then Y would be true*). For example, suppose I had just

held a fragile glass over a hard surface. Then the subjunctive implication *If I had dropped the glass just now, then the glass would be in pieces* is true. If, contrary to fact (hence the term *counterfactual*), I had just dropped the glass, the subjunctive implication proposes something else that—also contrary to what is actually the case—would have to be the case as well (namely, that the glass would be in pieces).

A material implication using the same antecedent and consequent is also true: *If I did drop the glass just now, then the glass is in pieces*. But (surprisingly enough) the material implication *If I did drop the glass just now, then the earth is flat* is also true. As noted just above, material implication merely asserts that we do not have a true antecedent together with a false consequent. In this last example, the antecedent and consequent are both false, so the material implication is true. With material implication, a false antecedent implies any consequent, true or false.

In contrast, the corresponding subjunctive assertion *If I had dropped the glass just now, then the earth would be flat* is not true (and thus, subjunctive implication, more than material implication, corresponds to our intuitive meaning of *if-then*, even though material implication is more familiar in mathematical logic). Subjunctive inference imposes a stricter condition than does material implication: a material implication follows from the corresponding subjunctive inference, but not vice versa. A subjunctive inference asserts that the antecedent somehow necessitates the consequent. If *X* does necessitate *Y*—meeting the subjunctive criterion—then we will not find *X* true and *Y* false—that is, we will find that the material-implication criterion is met. But when we do not find *X* true and *Y* false (for instance, when we find both false, as in the glass-dropping flat-earth example), it does not follow that *X* must necessitate *Y*. Saying just what the subjunctive necessitation consists of, though, turns out to be tricky and controversial, and is the crux of this chapter's discussion.

Of the four criteria listed above, the first three—evidential, subjunctive, and causal—often coincide. Say I take the action of crossing the street to achieve the goal of getting to the other side. (Let's construe that action as the initiation of a series of muscle contractions, not as the very passage across the street, so the goal's achievement doesn't just follow tautologically from the action's occurrence.) Knowing that I will cross informs you

(fairly reliably) that I will get to the other side, whereas knowing I will not cross informs you otherwise, fulfilling the evidential criterion. If I were to walk across the street, I would (very likely) get to the other side, but (very likely) not otherwise, fulfilling the subjunctive criterion. And finally, my walking across the street causes me to get to the other side, fulfilling the causal criterion. By any of those three criteria, there is a means–end link from the action of crossing to the goal of getting to the other side. Given that means–end link, and other things being equal, my desire to be on the other side rationally motivates my crossing.

In Newcomb's Problem, though, the criteria diverge. Taking just the opaque box, forfeiting the $1,000, is strong evidence that you obtain $1,000,000 in the opaque box, whereas taking both boxes is strong evidence that the opaque box is empty. But taking the transparent box or not has no causal influence on the content of the already-sealed opaque box. Thus, the evidential criterion says there is a means–end link from the action of taking just the opaque box, to the goal of obtaining $1,000,000 in the opaque box.[1] But the causal criterion says otherwise; if there is indeed a means–end link here, it is an acausal one. A similar divergence occurs in real-life Prisoner's Dilemma situations—requiring no fantastic science-fiction predictors—as argued in section 7.1. In fact, the causal and evidential criteria diverge even in some completely mundane and uncontroversial situations in which an action correlates with, but does not cause, a subsequent state, as discussed below in section 5.5.

The subjunctive criterion's verdict, meanwhile, seems ambiguous in Newcomb's Problem, in part because of the broad range of intuitions about the meaning of *would*. Indeed, in the loosest construal, subjunctive means–end links mimic evidential links: for example, if you were to take just the opaque box (as opposed to taking both boxes), then there would be—that is, would *have to* be (given that evidence)—$1,000,000 in the opaque box.[2] In a stricter construal, subjunctive links are just causal links: what would differ if you were to take just the opaque box—compared with your taking

1. Or at least, the most straightforward invocation of an evidential criterion endorses a means–end link here. Some authors (Eells 1982; Jeffrey 1983) have argued that a more nuanced evidentialist analysis says otherwise; see chapter 6, note 3.

2. Horgan (1981), for example, argues that counterfactual implications, properly construed for decision-making purposes, correspond to evidential links.

both boxes—is whatever taking just the opaque box *causes*, and nothing more.[3]

I argue in this chapter that there is a distinct, intermediate sense of *would*—narrower than correlation but broader than causation—that provides the correct subjunctive criterion for means–end relations—that is, a sense of *would* such that, to the extent that a goal would more likely obtain if a given action were taken than if not, the desirability of the goal rationally contributes to the motivation for taking the action. Call this the *choice-supporting* sense of *would*.

Even ignoring the divergence among the evidential, causal, and subjunctive criteria in Newcomb's Problem and the Prisoner's Dilemma (and a less dramatic divergence even in some mundane situations, as discussed in sec. 5.5), we need to establish that *some* such criterion is correct, rather than the fatalist criterion. That is, to establish that there are ever means–end links in a deterministic universe (or at any rate, in a deterministic enough universe—such as this one, I argue), we need to rebut the proposal that all choices are futile, given the inalterable preestablishment of all choices and outcomes in a deterministic universe.

Below, I briefly recapitulate (and agree with) a conventional argument that the evidential criterion is too lax—that it sometimes prescribes means–end relations where none exist (sec. 5.5). After outlining the case for the compatibility of choice and determinism (sec. 5.3), I argue that the causal criterion is too strict; I present a (comparatively) clear-cut example of an acausal means–end link (sec. 5.4). I then argue that the principles extracted from rebutting the fatalist criterion, and from considering a clear-cut acausal means–end link, lead to a subjunctive means–end criterion (sec. 5.6) that turns out to support the assertion of an acausal means–end link in Newcomb's Problem (chap. 6) and the Prisoner's Dilemma (chap. 7).

3. Pearl (2000) and Joyce (1999), for example, identify what follows counterfactually with what follows causally. Gibbard and Harper (1977), discussing Newcomb's Problem, refer to the evidential approach as *V-maximization* (maximizing the expected utility given an action, compared to the expected utility given some other action), in contrast with the counterfactual *U-maximization* approach, which they advocate. V-maximization computes conditional utility with respect to the conditional probability of an outcome given an action; U-maximization instead uses the probability that an outcome would occur if an action were to occur. Gibbard and Harper's U-maximization invokes considerations of physical law and causal independence, leading to a causal sense of *would*.

Before starting the game, let us pause to clarify the rules. Given a set of goals, what (putative) means–end criterion should you use as the basis for choosing actions, in order to best achieve those goals? The question is circular: it asks what means–end-recognizing policy's use is a means to best achieving your goals. We cannot deduce the answer unless we already have some basis for recognizing means–end relations.

Fortunately, this circularity need not be paralyzing. It is reminiscent of the status of inductive reasoning, which, as Hume noticed centuries ago, cannot be deductively supported. We may observe that induction has worked well in the past, but to therefore expect it to work in the future is circular. Nevertheless, we can easily see, in broad terms, how intelligent organisms would have evolved to use inductive reasoning, relying in part on some hardwired kernel of that reasoning; and similarly with means–end reasoning. Presumably, then, we perceive means–end links—at least in some routine situations—partly on the basis of some built-in criteria for their recognition; it's just what we're built to do (like breathing), whether or not we come up with a reason for it. We need not (and cannot) deduce the correctness of underlying criteria from first principles, nor need we construct an explicit formalization of those criteria—though the latter may at least be possible.

And if we do construct a correct explicit theory, gaining an explicit understanding of the means–end criteria that are built into our cognitive design, we can expect to extend our innate kernel of competence to deal properly with more subtle problems. Analogously, having a correct explicit formal theory of induction extends our ability to make accurate predictions, and thwarts the influence of incorrect theories we may fall prey to—such as the fallacy by which, after a series of coin tosses that came up tails, people naively expect an elevated probability of an ''overdue'' heads toss coming up next.

Just as a false overdue-heads theory can inspire suboptimal betting behavior, a false fatalist theory (for example) can inspire suboptimal goal pursuit. You probably know someone who would undoubtedly act to dodge an imminent collision, but who, on the other hand, declines to wear a seat belt on the grounds that when your time is up, there's nothing you can do about it. When consequences are sufficiently immediate and obvious, no one takes fatalist resignation seriously enough to behave accordingly. But when the consequences can only be appreciated with more-abstract

reflection (e.g., about a small probability of a large consequence), one's abstract conception of means–end relations becomes more influential, and a fatalist theory may indeed impede an appropriate action that a better theory would instead promote.

The game, then, is to use our means–end intuitions in clear-cut, routine situations to try to elicit explicitly the principles that drive those intuitions. Dennett (1980) calls such situations *intuition pumps*. We can then apply those explicit principles to trickier situations (like Newcomb's Problem and the Prisoner's Dilemma) where the intuitive verdict itself is vague or ambiguous (that application is analogous, say, to performing an explicit calculation of the statistical significance of a given sample, rather than just trusting what our inductive intuition says about it).

Crucially, the principles we extract from the clear-cut situations are not merely *descriptive* of our means–end intuitions in those situations. If all goes well, the principles turn out to be *metacircularly consistent*: the principles, applied to themselves, endorse those very principles. That is, the principles plug back into our means–end intuitions such that using those very principles (rather than some others, or none) will indeed strike us as a good idea, as a means to achieving our goals. Thus, the principles will be *prescriptive* as well as descriptive. They will tell us what means–end criteria we *should* use, where the should-ness is ultimately grounded in what our means–end intuitions prescribe in clear-cut exemplary situations, and what they prescribe if we explicitly contemplate the use of the means–end criteria themselves.

Accordingly, the structure of the following argument is to elicit means–end intuitions in (comparatively) mundane, uncontroversial situations—the choice machines of section 5.3, the hand-raising scenario in section 5.4, and the street-crossing scenario in section 5.5—and to sketch a means–end-recognizing mechanism to account for those intuitions. Having thus motivated the mechanism, I then recruit it to analyze Newcomb's Problem and the Prisoner's Dilemma.

5.3 Choice Machines

Consider a deterministic, artificial universe defined, say, by a computer program. Say the universe is superficially like our own, with three-dimensional space and various physical objects, including agents that have sensory

inputs, motor actions, and internal representations of various aspects of the state of the world. I make the case below that meaningful choice by such agents is possible, and that any real-world indeterminacy is therefore unnecessary (and also insufficient) to account for genuine choice.

Suppose the agents' cognitive design uses the prediction-value paradigm discussed in section 2.4.1. Thus, each agent's control system contains what I've called *schemas*—subjunctive assertions of the form *If conditions X apply* (the schema's *context*), *taking action Y would result in conditions Z* (the schema's *result*), in the choice-supporting sense of *would*, as discussed in the previous section. Each schema also specifies an estimated *reliability* (reliability is a consideration even with determinism, because an action's result can differ depending on conditions not mentioned in the context). Ignore for now the question of how the agent obtained these schemas; it just has them.

The context, action, and result conditions are each expressed as a predicate applied to the agent's current situation. The system also assigns quantitative *utilities*, positive or negative, to some conditions. The system can recognize when, according to the extant schemas, an action or a series of actions would (probably) result in a condition of positive or negative utility. This recognition influences the system toward or against taking those actions. The strength of the influence is proportional to the utility, and to the stated reliability of the assertions.

Despite the determinism of the agent and its universe, the following sorts of questions are answerable with regard to the agent:

> Why did it take that action? In pursuit of what goal was the action selected? Was that goal achieved? Would the goal have been achieved if the machine had taken this other action instead? The system includes the assertion that if the agent were to do *X*, then *Y* would (probably) occur; is that assertion true? The system does not include the assertion that if it were to do *P*, *Q* would probably occur; is that omitted assertion true? Would the system have taken some other action just now if it had included that assertion? Would it then have better achieved its goals?

Insofar as such questions are meaningful and answerable, the agent makes choices in at least the sense that the correctness of its actions with respect to its designated goals is analyzable. That is to say, there can be means–end connections between its actions and its goals: its taking an action for the

sake of a goal can make sense. And this is so despite the fact that everything that will happen—including every action taken and every goal achieved or not—is inalterably determined once the system starts up. Accordingly, I propose to call such an agent a *choice machine*.[4]

Seeing choice as a mechanical process that contemplates hypothetical actions and outcomes addresses some of fatalism's central challenges to deterministic choice:

- Why (in some cases) does it make sense for us to take an action for the sake of an already inalterably obtained goal? Answer: because (in those cases) if we didn't, the goal wouldn't obtain. This answer is perfectly compatible with the fact that the goal *had to* obtain. There is no contradiction, because the action had to obtain too.
- But why, then, does it make sense to contemplate alternative actions, and their results, and select from among them, if it is already determined which sole action is chosen? Similarly: because if not for that contemplation and selection, the preferred action wouldn't (necessarily) have been taken. This is compatible with the fact that it had to be taken. There is no contradiction, because the contemplation and selection—that is, the choice process—had to occur too.

In addition to making choices that are subject to teleological analyses, a choice machine comports with two intuitions that are plausibly among those that lead many people to conclude that their choices are not predetermined:

- Ordinarily, when you do X, you believe you did so only because, on the whole, at that moment, given the circumstances, you wanted to. If instead you had on the whole preferred at the time to do Y, you would have done that instead. Your choice is not bound by external constraints, except insofar as they exert influence through your preferences themselves (for instance, by attaching to an otherwise preferable action a consequence that you will regard as a penalty, e.g., if someone threatens to shoot you if you disobey).

4. As always, the question of terminology is not substantive. Whether or not we decide to use the label *choice*, we are discussing a machine that has goals and selects actions based on those goals, such that we can analyze whether its selection of a particular action in pursuit of a particular goal is sensible.

▪ If the universe is deterministic, then in principle someone could predict your choice in advance. But if someone did so and then told you the prediction, nothing would stop you from deliberately doing the opposite of whatever was predicted, thus making the prediction false. Its falsehood seemingly contradicts the statement that the prediction could be made in principle, thus in turn contradicting the premise that the universe is deterministic.

The first of these intuitions is as true of an artificial choice machine as of a human being: the machine's preferences and machinery do indeed control its actions, and if its preferences had dictated a different action, then its action would have been correspondingly different. Of course, its preferences and knowledge are in turn caused by past events beyond its control; it cannot just choose its current preferences as it can its current actions. But neither can we.

The second intuition also applies to artificial choice machines as well as to human beings. The determinism of the choice machine's universe indeed implies that its choices are predictable in principle. And its world might easily be so constructed that an entity embodied in that world could carry out such a prediction—say, by taking a snapshot of a sufficient portion of the state of the world (including the choice machine) and then applying the deterministic laws of the world to calculate in advance how that portion of the world will evolve.

It is also true that the machine's preferences could be such that if the mechanism were told which action it is about to choose on a given occasion, it would be sure to choose a different action instead. But this ability to thwart a prediction does not contradict the determinism of the machine's world (as discussed, e.g., in MacKay 1960 and Dennett 1984). If an entity embodied in the choice machine's world is going to predict the machine's choice and then tell the machine the prediction before the choice is made, then the predictor cannot necessarily carry out the prediction in the first place. For to carry out the requisite simulation, the predictor must, among other things, simulate the choice machine's being told the prediction. To do that, the simulation must specify which prediction will be conveyed. But it doesn't know which prediction will be conveyed until the simulation is complete.

The predictor might, of course, carry out two separate simulations (assuming a binary choice of available actions), each presupposing a different

prediction to be conveyed to the simulated choice machine in that simulation. (Such a tactic figures in the discussion below in sec. 6.2 of a Newcomb's Problem variation in which both boxes are transparent.) If either simulation then predicts a chosen action that accords with the prediction conveyed in that simulation, then that prediction can be made accurately, and then conveyed to the choice machine for real. But if both simulations show the conveyed prediction being thwarted—as will be the case if the choice machine's preferences are such that it always acts to contradict the conveyed prediction—then the prediction cannot be accurately made and conveyed. Despite the determinism of the choice machine's universe, including the choice machine itself, the machine, like us, can easily choose to thwart a prediction that is conveyed to it in advance of the choice.

Thus choice, in the sense just delineated, is a mechanical process compatible with determinism: choice is a process of examining assertions about what would be the case if this or that action were taken, and then selecting an action according to a preference about what would be the case. The objection *The agent didn't really make a choice, because the outcome was already predetermined* is as much a non sequitur as the objection *The motor didn't really exert force, because the outcome was already predetermined*. (Or as Dennett puts it in *Elbow Room*, wouldn't a predetermined thunderstorm still be a real thunderstorm?) Both choice making and motor spinning are particular kinds of mechanical processes. In neither case does the predetermination of the outcome imply that the process didn't really take place.[5]

Thus, a fully deterministic universe would not preclude genuine choice. But some degree of real-world indeterminacy would also be compatible with choice:

- Indeterminacy of some external events is the same, for most ordinary (nonquantum) purposes, as an agent's limited-knowledge perspective—a perspective that makes an outcome seemingly random to the extent that the outcome depends on definite but unknown conditions. But excessive external indeterminacy (or excessive unpredictability due to limited knowl-

5. The term *free will* scarcely appears herein. As typically used, that term connotes something like *meaningful choice, as opposed to determinism*. Since I am defending meaningful choice, even given determinism, I avoid the f-w word and refer simply to the choice process.

edge) would make choice futile: if just about anything could happen (or could happen as far as you know) regardless of what you were to do, there'd be no point in making any particular choice instead of some other.

▪ Likewise, some indeterminacy within the choice process itself is harmless, say as a tiebreaker among choices of comparable utility. But too much internal indeterminacy would again make choice futile: if, say, seeing an oncoming car, and strongly preferring not to be struck, did not determine reliably that you'd choose to step out of the way, then choice would be of little use.

Thus, moderate indeterminacy is compatible with choice, but not necessary for it. Profligate indeterminacy—contrary to a popular intuition—is subversive of choice, rather than supportive of it.[6]

Accordingly, in the examples that follow, I presume a predictable universe, either by idealizing our own universe as such (both by presuming deterministic physics, and by treating some minuscule uncertainties—uncertain because of limited knowledge—as having zero rather than minuscule probability) or by considering an artificial, deterministic universe populated with choice machines. The lessons extracted apply to the real world insofar as real-world choice is not dependent on whatever moderate indeterminacy may exist.

5.4 Acausal Means–End Links: Choosing Past States

The conclusion so far is that—even given determinism—the fatalist criterion of means–end links is incorrect in its insistence that no such links exist. I next consider the other proposed criteria enumerated in section 5.2, starting in this section with the causal criterion.

6. Even under Everett's relative-state interpretation (which chap. 4 above argues for), quantum mechanics, though technically deterministic (in that quantum amplitude flows deterministically through configuration space), still has the property that a given classical state (in some configuration-space branch) is followed by divergent classical successor states (in subsequent branches). This divergence resembles nondeterminism as far as choice is concerned: either way, the present is compatible with multiple futures, in some sense or other. Whether or not the multiplicity is genuinely nondeterministic, the current argument holds: some degree of multiplicity of futures is compatible with choice (though not required), but excessive multiplicity would undermine choice.

Determinism does not imply fatalism, because inalterability does not imply futility. As just discussed, it can make sense for an agent to (be designed to) choose an action for the sake of what would be the case if the action were taken (in some appropriate, choice-supporting sense of *would*), even though all events (including the achievement or nonachievement of any goals) are already inalterably determined.

Must an action *cause* a goal's achievement in order for there to be a means–end relation? A compelling reason to think so is that in the absence of a causal link, the goal's achievement is inalterable by the action. But as just discussed, any goal's achievement is inalterable anyway, given determinism; still, it can make sense to act for the sake of a goal. Such is sometimes the case, I claim, even for a goal that is not caused by an action.

Consider a simple example, presuming that our universe turns out to be deterministic (or consider a similar example set in a deterministic alternative universe). Define the predicate *P* to be true of the total state of the universe at a given moment if and only if the successor state one billion years thence—that is, the state defined by applying the (correct) laws of physics to the given state to predict the new state one billion years later—shows me with my right hand raised. Suppose, on a whim, I would like the state of the universe one billion years ago to have been such that the predicate *P* is true of that state. I need only raise my right hand now and voilà, it was so.

Of course, I did not change what the distant-past state of the universe had been. The past is what it is and can never be changed. Furthermore, I have no causal influence over the past. Nonetheless, physical law (if deterministic) necessitates that if I do in fact raise my right hand, *P* is in fact true of the state of the universe a billion years ago; or if I do lower it, *P* is false of that past state. Suppose, despite my wanting *P* to have been true a billion years ago, I forgo raising my hand due to a belief that it would be futile to act for the sake of something past and inalterable. In that case, as always with fatalist resignation, I would be needlessly forfeiting an opportunity for my goal to be achieved.

The set of universe-states in which I do raise my hand is necessarily coextensive with the set of universe-states for which, in the state a billion years prior, *P* was in fact true (in other words, those are two different descriptions of the same set of universe-states). I thus have *exactly* as much choice about that particular aspect of the past, despite its inalterability, as I have about

whether to raise my hand now—despite the inalterability of that too in a deterministic universe. And, as argued in the previous section, that much choice is choice enough.

There can also be an acausal means–end link in the presence of some uncertainty. Suppose, for instance, that the world's physical laws are nearly deterministic, but with some minute chance (say 10^{-1000}) that the laws by which a distant-past state ordinarily leads to my raising my hand (or to the absence of that event) have an exception in a particular instance. There is still an (acausal) means–end link from my raising my hand, to the past state such that I would (almost certainly) raise my hand. However, that link is now (negligibly) attenuated by the minuscule chance that a different past state—one that is *almost* certainly not a hand-raising precursor—led to hand raising on this occasion. If, in the fully deterministic case, I would raise my hand for the sake of the past-predicate goal, it would make no sense to do otherwise due to a mere 10^{-1000} chance of an exceptional physical event.

A clarification is in order about the meaning of *cause*. One might maintain that a means–end relation is inherently causal by the very definition of cause: if you successfully choose whether something is the case, then tautologically you thereby cause it to be the case or not.

But this issue is merely a matter of terminology. I wish to distinguish two concepts: the concept of a chain of influence by one partial state of the universe on another and then another, according to physical laws (roughly speaking, one particle bumps into another, which then bumps into another ...); and the concept of a means–end relation between an action and a goal state, such that if the state is desired the mechanism should take the action, other things being equal. I use *cause* to designate the first of these concepts. Regardless of terminological conventions, though, the substantive question here is whether the first of these is necessary for the second, or whether instead a subjunctive link can constitute a means–end link even in the absence of a physical-influence link from action to goal—that is, in the absence of a causal link, as I use that term.

Needless to say, when I speak of acausal means–end links, I am not proposing some sort of acausal "force" or process in the universe, in addition to a causal force or process. On the contrary, all the events in the spacetime of a deterministic universe (or at least the sort of deterministic universe contemplated here) are specifiable by the initial state and the (causal)

physical laws that say how each state determines the immediately temporally adjacent states; there is no room left over for any other principles to specify what happens (except, of course, insofar as those principles merely recapitulate or summarize or approximate some aspect of the underlying physical laws). Both causal and acausal means–end relations are just that: relations. That is, they are particular abstractions among the (actual or hypothetical) events being related. The relations' ontological status, on this account, is the same as the ontological status of being correlated, which (rather than being a force or process) is likewise just another abstract relation among events—albeit an easier one to formalize and analyze. The question at hand is: *what sort* of relation among events constitutes a means–end relation?

The remainder of this chapter looks more closely at what relations are means–end relations. So far, the above discussion of the past-predicate example argues that means–end relations can indeed be acausal. Exhibiting an example, however trivial, of an acausal means–end link suffices to rebut the contention that means–end links are necessarily causal, and clears the way for more-interesting examples. I argue next (sec. 5.5) that means–end links, although sometimes acausal, are not merely evidential—there is more to a means–end link than that the action's occurrence gives evidence (however strong) of the goal's occurrence. Instead, I maintain, means–end links are *subjunctive* relations (sec. 5.6). Sections 6.1 through 6.3 discuss the acausal subjunctive means–end links that arise in a few versions of Newcomb's Problem. Then, section 7.1 ties Newcomb's Problem to the Prisoner's Dilemma, and examines the acausal means–end links in that scenario as well.

5.5 Street-Crossing Scenario: Avoiding Evidentialist Excess

The means–end link to predicate *P* above is inconsistent with a causal criterion for means–end links, but is consistent with an evidential criterion: the conditional probability of *P* being true of the distant past given that I raise my hand (i.e., a probability of 1) is indeed higher than the probability of its being true given that I do not raise my hand (i.e., a probability of 0).

But whereas a causal criterion is too strict—incorrectly excluding what I have argued are some genuine instances of means–end links—an evidential criterion is too lax, sometimes wrongly designating means–end links where

none exist. This section examines the excesses of evidentialism, in part to jump on the bandwagon of its critics (e.g., Pearl 2000), but also because the subjunctive alternative I propose in the following sections can be seen as a modification of evidentialism, adopting technical fixes for specific problems discussed here, while keeping much of the evidentialist framework.

To illustrate where evidentialism goes astray, consider the following scenario. I stand on a street corner, wanting to be on the other side of the street. I have a clear view of any oncoming traffic and have looked carefully. I strongly prefer not to be struck by traffic, so I would not cross the street at a moment when I see dense traffic just a few meters away, speeding toward the intersection. Let us make the idealizing assumption that the probability of my crossing under such circumstances, given my preference not to be struck, is zero; and assume I know it. Thus, we ignore the minute possibility that, say, some cosmic rays will disrupt my neurons such that I knowingly cross dangerously now despite my preference not to and despite my competence, having looked carefully with a clear view, to act accordingly.

To evaluate whether a particular action would constitute a means to achieving a particular goal state, evidentialism compares the conditional probability of the state given that the action is taken, and the conditional probability of the state given that the action is not taken. But if I already know I will not cross in the presence of clearly seen dangerous traffic—if the probability of my doing so is zero—then any probability conditioned on my crossing under those circumstances is simply undefined (because the defining expression is a quotient with a zero denominator; the conditional probability $\Pr(A|B)$—the probability of A given B—equals $\Pr(AB)/\Pr(B)$; roughly speaking, that quotient of probabilities tells us what proportion of B events are also A events, and hence are AB events). Accordingly, an evidentialist might hold out for the cosmic rays and refuse to accede to the zero-probability idealization.[7]

But let us recast the scenario in an artificial world where a choice machine (with street-crossing knowledge and preferences similar to mine) is

7. Alternatively, an evidentialist might side with, e.g., Hájek (2003) and endorse a nonstandard analysis of conditional probability according to which some probabilities are meaningful even if conditioned on a zero-probability event. But then, as discussed below, ambiguity arises as to how to assign the value of such a conditional probability.

waiting to cross the street, but sees dangerous traffic. Suppose the probability of its crossing right now, given its preferences and sensory inputs, is literally zero (say the world's physics flatly precludes interference by cosmic rays, or any other hardware failure of the choice machinery), and is correctly represented as such by the choice machine itself.

Despite that certainty, the machine must still choose, on some basis, whether or not to cross the street. As discussed in section 5.3, the inalterability of the forthcoming choice process—in this case, a process with an already known outcome (namely, not crossing)—does not imply that no choice is being made. On the contrary, the choice process—comparing what would happen if this or that action were taken, and initiating the action for which what would happen is preferred—operates as always, and crucially so. For it is that very process—the choice machine's anticipation of what the consequences would be, and its initiating an action on the basis of that anticipation—that *makes* it impossible for the machine to cross now, in the presence of clearly seen dangerous traffic.

The artificial-world thought experiment shows that a nonzero probability of crossing in the presence of dangerous traffic is not necessary for there to be (by some intuitively clear but yet-unexplicated criterion) the usual means–end link, in the presence of dangerous traffic, from the action of crossing, to the (negative-utility) outcome of being struck by traffic (or equivalently, from the action of not crossing, to the goal of not being struck). Thus, even if my crossing dangerously—despite what I see and what I prefer—does have a slightly nonzero probability in the real world, that negligible possibility—along with the cosmic rays, or whatever, that bear it—is irrelevant to why I should not cross the street when the traffic is dangerous. The zero-probability idealization thus does not forfeit any necessary explanation for why I should not cross.

A preliminary problem, then, with evidentialist means–end links is that they make it difficult to say what outcome an action would lead to when known circumstances preclude the action—even if, as in the present example, they preclude the action *because* of what the action would lead to (i.e., I will not cross now because of what I know would then occur).

Potentially mitigating this problem, we might allow ourselves some latitude about what probabilities we bring to bear, while still invoking an evidentialist approach. If we condition on only *some* relevant aspects of the situation, then the requisite conditional probabilities can be defined, as ela-

borated below. But unfortunately, it turns out that that approach can yield an absurdly wrong answer, as follows.

In preparation, let us define a few assertions:

S (Street) An agent stands at a street corner ready to cross the street.

D (Disposition) The agent's desire, competence, and view of the street are such that the agent will cross now only if there is no dangerous onrushing traffic.

C (Cross) The agent crosses the street now (i.e., the agent initiates a particular series of muscle contractions intended to move the agent forward).

T_1 (Traffic$_1$) Dense, dangerous traffic is just about to speed through the intersection.

T_2 (Traffic$_2$) Dense, dangerous traffic arrives at the intersection a second later (i.e., just as crossing begins, if that action is now taken).

Each of these indexical assertions is a predicate that applies to a given situation. Making the deterministic idealizations above (or recasting the scenario in a suitably deterministic artificial world), we can then express probabilities such as

$\Pr(T_2|SDC) = 0$

$\Pr(T_2|SD{\sim}C) = 0.3$ (say).[8]

That is, among situations that satisfy *SDC*, T_2 is never satisfied—dangerous traffic is never seen to arrive just as I cross (because of my desire and competence not to cross dangerously). But among situations that satisfy $SD{\sim}C$, T_2 is often satisfied; such traffic often does arrive when I do not cross. The above probabilities express how the likelihood of the arrival of dangerous traffic varies as a function of whether or not I cross, given the street-crossing situation and my desire and competence to cross safely.

Assume that I know that *S*, *D*, and T_1 are true right now. By the conditional probabilities above, my taking the action of crossing the street—given my disposition regarding safety—is perfect evidence of the absence of dangerous traffic as I cross. And indeed, a third party with no view of the road itself would be justified to bet on the absence of dangerous traffic,

8. A more careful notation would specify that the predicates gathered in the probability expression must all be applied to the same situation. But that's obvious enough to be left implicit, here and throughout.

just knowing my disposition and that I chose to cross. As a passive prediction, the conditional probability $\Pr(T_2|SDC) = 0$ is impeccable—it correctly assesses (as 0) the probability of finding the arrival of dangerous traffic at the moment that a safely disposed agent (as defined above) is crossing.

Crucially, though, that evidential relation obviously does not suffice to establish a means–end link from the action of street-crossing to the goal that there be no dangerous traffic. Even if I prefer the absence of dangerous traffic (say, so I can continue across the street safely), it would be foolhardy and bizarre of me to cross the street now—as the dangerous traffic plainly approaches—for the sake of achieving that traffic's absence—that is, to misconstrue the evidential link as a means–end link.[9]

The conditional probability $\Pr(T_2|SDC) = 0$ predicts what one finds regarding T_2 whenever one actually finds *SDC*. But it does not necessarily predict what one *would* find now regarding T_2 if one were to bring about C now given that *SD* is actually the case now. Thus, the second problem with an evidentialist criterion of means–end links is that, unsurprisingly, mere evidence—mere correlation between an action and a goal state—does not necessarily amount to a means–end link.

Potentially mitigating this second problem, we might appeal to other relevant probabilities, such as

$$\Pr(T_2|ST_1C) = 1.$$

That is, the oncoming traffic (T_1) does not just vanish as it reaches the intersection, even if crossing occurs—the traffic still arrives (T_2). Empirical support for this probability can be derived, for instance, from observing the plight of agents who are *not* of safe disposition.

The prediction about T_2-if-C expressed by $\Pr(T_2|SDC) = 0$ is contradicted by the prediction expressed by $\Pr(T_2|ST_1C) = 1$. Given the agent's safe dis-

9. A similar point is often made, as by Nozick (1969), in terms of an imaginary scenario in which we discover that smoking does not cause cancer; rather, a gene that predisposes people to smoke also independently predisposes them to cancer. Then, it would be irrational to avoid smoking in order to avoid cancer; there is a correlation, but no causal link and no means–end link, from not smoking, to avoiding cancer.

The smoker's scenario is similar to the street-crossing scenario: the gene (like T_1) is a common causal influence on both a choice (smoking, or $\sim C$) and on a second effect that is thereby correlated with the choice (cancer, or T_2); but the choice does not cause the second effect, nor does making the alternative choice serve as a means to avoiding the second effect, despite its giving evidence of the second effect's absence.

position, crossing is perfect evidence for the absence of arriving dangerous traffic T_2. But given oncoming dangerous traffic T_1, crossing is no evidence at all for that absence (T_1 is said to *screen off* T_2 from C, meaning that $\Pr(T_2|ST_1) = \Pr(T_2|ST_1C)$—i.e., if we condition on T_1, then further conditioning on C does not change the probability of T_2).

Since we know both D and T_1, we face an ambiguity about which of those predicates to condition on. We would like to use both; in general, we should condition on information as specific as is available and relevant. But as noted above, a probability conditioned on SDT_1C is undefined, under the assumption that $\Pr(SDT_1C) = 0$.

Regarding what *would* happen now if I were to cross, $\Pr(T_2|ST_1C) = 1$ obviously expresses a more intuitively plausible prediction here than does $\Pr(T_2|SDC) = 0$—that is, we know the oncoming traffic wouldn't vanish if I were to cross now. But on what do we base that knowledge?

True, one might observe that given the dangerous traffic T_1 (which is in fact present), if I do cross now, it cannot be true that I have D. Hence, one might argue that when ascertaining what C is evidence for, we should not condition on D, because D itself logically depends on C. But symmetrically, one might observe that given my disposition D (which is in fact present), if I do cross now, it cannot be true that dangerous traffic T_1 is present. Hence, T_1 apparently depends on C, so we seemingly should not condition on T_1. It is only because we (somehow) already know intuitively that C is not a means to $\sim T_2$ that we know which predicate, D or T_1, to condition on—not vice versa. (Sec. 6.2.1 below revisits the tricky question of whether to condition on something that in some sense depends on x, when we ascertain what result follows from x.)

Thus, the evidentialist means–end criterion is at best ambiguous. It endorses both an absurd means–end link (in that safely disposed crossing is perfect evidence for the nonarrival of dangerous traffic) and a correct, contradicting assertion (that oncoming traffic does not just vanish when an agent—presumably not a safely disposed one—steps in front of it), depending on which of two actually true assertions we condition on (we cannot condition on both together, for the reasons discussed above). But the evidentialist criterion provides no principled way to resolve the ambiguity.

We might marshal other evidence, too, for each of the opposing predictions—appealing, for instance, to general principles by which we

could anticipate either of the conflicting conditional probabilities, even in the absence of direct empirical evidence (even if you've never watched anyone cross a street, you could derive those conditional probabilities from the problem description and from more-general knowledge about the world). But piling on more evidence to bolster each of the opposing predictions would only compound the contradiction (unless some piece of evidence just quantitatively overwhelmed the contrary evidence, in a Bayesian manner; but when the conflicting evidence already involves probabilities of 0 or 1, quantitative overpowering is out of reach). We need, instead, a way to decide that the absurd means–end link should not even be in contention, in order to conclude (as we should) that the probability is essentially zero that the oncoming traffic would vanish if I were to cross now.

To that end, one might propose an ad hoc rule that when we appraise a putative means–end link by computing the conditional probability of an outcome given an agent's action, we should not condition in part on an agent's competence or desires.[10] That rule would obviate the problem in this particular example—conditioning on *D* would just be outlawed. Indeed, some simpler organisms that (presumably) lack the ability to reflect on their own competence and desires might thereby circumvent such difficulties.

We humans, though, can and do base our expectations and plans in part on such self-assessments. For example, whether you drive in a snowstorm at night may depend on whether you think you're awake enough to do so safely. Similarly, whether you embark on a challenging project may depend on whether you expect to remain motivated long enough to finish it. Sacrificing the ability to condition on one's competence or desires would be a high price to pay to deal with the evidentialist problem.

Moreover, such a sacrifice would not solve the problem anyway, at least in principle. Suppose someone has made a detailed, accurate copy of an agent's cognitive state. The agent could then condition on the state of that copy instead, circumventing the proposed ad hoc rule. One might try

10. Christopher Taylor (personal communication) suggested such a rule. Others, contemplating different scenarios (see n. 12), propose, on the contrary, that we *must* condition on the agent's dispositions and decisions (which in this case yields the wrong answer). Devising rules for means–end recognition is tricky in part because of the temptation to postulate a rule that gives the right answer in the scenario under consideration, without noticing that it gives the wrong answer in other, equally fundamental examples.

to plug this loophole by also forbidding conditioning on states that are even strongly *correlated* with the agent's cognitive state. But then I could not even condition on T_1 in the street-crossing scenario—that is, I could not even take account of the oncoming traffic in assessing what would happen if I were to cross now—since my inclination to cross correlates strongly with the absence of oncoming dangerous traffic.

Here again, an evidentialist might chafe at the zero-probability idealization, and insist on invoking minuscule but nonzero probabilities in order to rebut the false means–end link from C to $\sim T_2$. The evidentialist is then able to define the conditional probability

$$\Pr(T_2|SDT_1C) = 1,$$

conditioning on both T_1 and D. This is the probability that the oncoming traffic does not vanish, given the hypothesis that despite an actual desire and competence to the contrary (and having looked carefully with a clear view, etc.), an agent does cross now in the presence of the dangerous traffic. Intuitively, we know that that conditional probability is 1 (or nearly so) if the conjunction SDT_1C is not quite impossible.

But in the absence of any actual instances of the all-but-impossible event of an agent's crossing in the path of clearly seen dangerous traffic despite the agent's contrary desire and competence D, on what basis can the evidentialist assign a conditional probability to the continued presence of traffic in that hypothetical situation? The same dilemma arises as with the zero-probability idealization: how do I know whether the oncoming traffic would proceed now as it is always in fact observed to do (i.e., by not suddenly vanishing), or whether instead my act of safely disposed crossing would proceed now as it is always in fact observed to (i.e., safely and successfully, with no dangerous traffic passing as I cross)? Of course it is intuitively obvious which would occur, but how is that intuition implemented?[11] On what basis can we conclude that $\Pr(T_2|SDT_1C)$ equals $\Pr(T_2|ST_1C)$ rather than equaling $\Pr(T_2|SDC)$? (An ad hoc rule against conditioning here on the agent's disposition is undesirable and inadequate, for the same reasons as noted above.)

11. The overwhelming intuitive obviousness of the right answer (as to which outcome would occur) can obscure the fact that there is even a problem here to be solved. Envisioning the design of a machine that can figure out that answer helps bring the problem into focus.

One might just invoke a raw *subjective* (not to be confused with *subjunctive*) conditional probability (as to what transpires given the almost impossible and never actually observed hypothetical event SDT_1C), and then assign means–end links according to the subjective probability. That tactic, however, does not explain how to ascertain the means–end link here at all, but rather just passes the buck to whatever homunculus generates the subjective conditional-probability intuition. The putative explanation thus circularly presupposes that something somehow has already solved the problem. The explanation I seek addresses *how* the intuition here, corresponding to the subjective conditional probability $\Pr(T_2|SDT_1C) = 1$, could reasonably be arrived at by our cognitive machinery, given the contradictory evidence (as to T_2-if-C) offered by actual SDC situations (where T_2 is never true) and by actual ST_1 situations (where T_2 is always true), and the absence of actual SDT_1C situations (even if they are not quite impossible).

Thus, even putting aside the zero-probability idealization, evidentialism is stuck with an ambiguity about which relevant information to condition on when deciding whether or not the traffic would just vanish if I were to cross.[12]

12. Horgan (1981) defends evidentialism against Nozick's smoker's analysis (n. 9 above) by appeal to *screening off* by one's knowledge of one's inclination prior to acting: conditioned on a smoker's knowledge of her desire to smoke (which desire is what we imagine mediates the gene's influence on smoking), the act of smoking itself contributes no *further* evidence as to the gene or the propensity for cancer. Similarly, Eells (1982) and Jeffrey (1983) argue that in smokerlike scenarios, knowledge of the putative result is screened off from knowledge of the action by the knowledge of one's decision just prior to acting (assuming at least an arbitrarily small probability that the decision and action are discrepant, to allow the requisite conditional probabilities to be defined).

But in the present example, T_1 already serves much the same screening-off function as does the foreknown decision, due to its strong correlation with that decision. And if the analysis were recast, explicitly conditioning also on the already-known decision not to cross ($\sim C_D$), and thus comparing $\Pr(T_2|SDT_1{\sim}C_DC)$ to $\Pr(T_2|SDT_1{\sim}C_D{\sim}C)$, the above problems would remain: 1) the analysis now requires a nonzero probability of ${\sim}C_DC$ and of SDT_1C, but a thought experiment set in an artificial world where those probabilities are zero shows that a coherent choice is still possible; and 2) even with those probabilities slightly nonzero, the agent, never having actually encountered ${\sim}C_DC$ or SDT_1C, has no way to ascertain whether $\Pr(T_2|SDT_1{\sim}C_DC)$ equals $\Pr(T_2|SDT_1{\sim}C_D) = 1$ or instead equals $\Pr(T_2|SDC) = 0$, unless the agent has already somehow solved the very problem under discussion.

A causalist, of course, is not at risk for the present dilemma about which evidential relation (the one conditioned on T_1, or the one conditioned on D) to use in assessing whether there is a means–end relation. Because there is no causal link from my crossing the street, to the absence of dangerous traffic—there is a causal link between the two, but it points the other way—the causalist acknowledges no means–end connection.[13] But the causal criterion is too strict, subjecting the causalist to fatalist resignation concerning some readily achievable goals, as in section 5.4's past-predicate example (where the causal link also points in the opposite direction of the means–end link). An agent that perceives a means–end link in the past-predicate scenario thereby achieves its past-predicate goals (and other, less whimsical goals, discussed below) better than does an agent using only a causal means–end criterion. But an agent that (mis)perceives a means–end link from crossing to no-dangerous-traffic does not thereby achieve its no-dangerous-traffic goal.[14]

If a causal link is unnecessary for there to be a means–end link, and an evidential link is insufficient, how then is a means–end link to be recognized? My proposal is that an evidential link—defined as above in terms of contrasting probabilities conditioned on a contemplated action's occurrence or nonoccurrence—works well as a presumed means–end link, but the presumption needs to be defeasible in some cases, as discussed in the following section.

Informally, the refuting intuition is easily stated in the street-crossing scenario: yes, given my safe disposition, there's certain to be no dangerous

13. A causalist does, however, need to be concerned about what conditional probabilities to use to recognize the relevant causal relations in the first place, as addressed, e.g., by Pearl (2000).

14. Here I invoke metacircular consistency as discussed in section 5.2: using the past-predicate means–end construal is a means to the goal of achieving (some) past-predicate states, but using the no-traffic means–end (mis)construal is not an effective means to achieving the no-traffic goal. Metacircular consistency is not definitive, because it is circular: to apply it, we need to know by what criterion (causal, subjunctive, evidential, or whatever) we can say that using a given means–end construal policy is a means to achieving one's goals. But by *any* of those three criteria, it is counterproductive to depend on a putative means–end link from crossing to no-dangerous-traffic; by any of those criteria, if we were designing an agent, designing it to invoke that putative means–end link would not be a means to ensuring that the agent best pursues its goals. So metacircular consistency does give us some purchase on the problem.

traffic arriving whenever I do cross the street. But that's only because that correlation between safely disposed crossing and no-dangerous-traffic is never tested (by crossing the street) in precisely the circumstance where the traffic would be found to arrive (namely, when it is already almost there). In the next section, I attempt to present more formally the foregoing intuition about how an evidential link can be superseded, countering the preliminary presumption that it corresponds to a means–end link. I argue that an evidential link, coupled with a way to selectively supersede it, amounts to a subjunctive link, which is correctly construed as a means–end link.

5.6 Subjunctive Means–End Recognition

The choice machine in section 5.3 uses subjunctive assertions—means–end links—of unspecified origin. But consider now an *analytical* choice machine—one that assesses for itself the validity of proposed means–end connections. Given the circularity noted in section 5.2, at least some kernel of the machine's means–end recognition must be built in; otherwise, even if the machine could reason well enough to figure out that using a given means–end criterion would be advantageous, it would not thereby be influenced to use that criterion!

This section sketches aspects of the design of an analytical choice mechanism. The design effectively defines a proposed subjunctive criterion for means–end links, intermediate between evidential and causal criteria (evidential and causal criteria, as the preceding two sections argued, are respectively too lax and too strict). I begin by outlining an essentially evidentialist relation that a choice machine might use. Then, I introduce a further condition to try to limit the relations to choice-supporting subjunctive ones, that is, means–end links.

5.6.1 Choice Machines and Schemas

Consider predicates such as the indexical assertions (i.e., assertions about a pointed-to situation) defined in section 5.5, defined now from the point of view of an analytical choice machine. As in section 2.4.1 above, let us use the notation

$C_1 \ldots C_i : A_1 \ldots A_j \rightarrow R_1 \ldots R_k \ (r)$

to represent a *schema*. The schema asserts that if context predicates $C_1 \ldots C_i$ are satisfied in the current situation (i.e., that they are true when applied to the current situation), and if action conditions $A_1 \ldots A_j$ are also satisfied in that situation, there is probability r (the schema's *reliability*) that conditions $R_1 \ldots R_k$ are satisfied; that is,

$$\Pr(R_1 \ldots R_k \mid C_1 \ldots C_i \; A_1 \ldots A_j) = r.$$

Let us assume that the choice machine treats probabilities as frequencies among actual situations. Then, the machine can empirically verify or adjust a schema's specified reliability, assuming the machine encounters a large enough sample of relevant situations.[15] When the schema notation omits a designated reliability, assume the reliability is 1.

Schema notation introduces a distinction between context and action, whereas the underlying conditional probability simply conditions on the conjunction of the context and the action. The context–action distinction acquires operational significance by virtue of the machinery proposed below, which reinterprets schemas to designate more than just conditional probabilities.

I do not address here how a choice machine might propose particular schemas for consideration in the first place (out of the exponentially many combinatorial possibilities), how it might define the predicates in terms of which its schemas are expressed, or how it might ascertain whether a given predicate is currently satisfied (but my book *Made-Up Minds* offers some suggestions). What concerns me here instead is: *given* that an agent somehow amasses knowledge about how the world *is*—evaluating the current truth of various predicates, and constructing schemas that tabulate some correlations among the predicated states' actual occurrences—how can the agent get from there to knowing how the world *would* be (in the choice-supporting sense) if this or that action were now taken? Given a set of predicates and schemas, how does the agent's machinery then recognize means–end links? Accordingly, in this discussion, I just postulate the presence of whatever (accurately maintained) predicates and schemas are needed to illustrate the abilities and vulnerabilities of the proposed means–end-recognizing machinery.

15. Alternatively, a choice machine might ascertain some schema probabilities simply by being told what they are, or by other techniques that are indirectly grounded in observations. But for present purposes, a presumption of direct empirical sampling will suffice.

Extending the notions introduced in section 2.4.1, a given schema is said to be currently *applicable* when its context conditions are all satisfied in the current situation, and when the schema is not overridden in one of several ways discussed below. An otherwise-applicable given schema is subject to an *exception override* if another currently applicable schema asserts a different probability for the same result given the same action, conditioned on context predicates that designate a strictly more specific condition than the given schema's context (i.e., the given schema's context conditions are always satisfied when the other's are). An exception-overridden schema is considered currently inapplicable; the overriding schema's probability is thereby asserted in place of the overridden schema's probability. Thus if the choice machine has, say, schemas

$C_1{:}A_1{\rightarrow}R_1$ (0.90)

$C_1C_2{:}A_1{\rightarrow}R_1$ (0.14) (or equivalently, $C_1C_2{:}A_1{\rightarrow}{\sim}R_1$ (0.86)),

and C_1, C_2 are both satisfied now, the first schema is currently exception-overridden by the second schema, which asserts just a 0.14 probability of R_1 if A_1. But if the choice machine did not have the second schema, the first schema would now assert a 0.90 probability of R_1 if A_1. Thus, the choice machinery effectively makes a preliminary presumption that the probability expressed by a schema is conditionally independent of aspects of the world not designated in the schema's context. But that presumption can be overridden by another schema (with its own empirical support) that applies under more-specific conditions.

Assume that if the choice machine has schema $C{:}A{\rightarrow}R$ (x), it also has a *complementary* schema $C{:}{\sim}A{\rightarrow}R$ (y) (or equivalently, $C{:}{\sim}A{\rightarrow}{\sim}R$ (1-y)). That is, the machinery keeps track of the result's probability both with and without the specified action (given the context).[16]

Arbitrary conditions (not just personal motor events) can appear in the action part of a schema (and in the context or result). By the *composition* of schemas (discussed just below), the choice machinery can sometimes use other schemas to bring about the satisfaction of a given schema's designated action conditions.

16. In *Made-Up Minds*, a given schema keeps track of both probabilities, combining what I here call two complementary schemas. The difference is just a change of terminology.

Each predicate has a *utility* attributed to it (positive, negative, or zero), allowing some predicates to serve as goals. If $C{:}A{\rightarrow}R\ (x)$ and $C{:}{\sim}A{\rightarrow}R\ (y)$ are currently applicable, and R has utility u, the pair of schemas currently attribute to action A the attenuated utility $u(x\text{-}y)$ with respect to result R—a conventional expected-utility calculation.[17]

For each designated action, the machinery keeps track, from moment to moment, of the utilities currently attributed to that action by the then-applicable schemas that use that action. The action to which is currently assigned the greatest net attenuated utility is selected for activation. Thus, the mechanism is influenced to take an action from which there is a link—via a currently applicable pair of complementary schemas—to a state of positive utility (while avoiding actions that link to a state of negative utility).

By virtue of the foregoing provisions, the choice machinery treats schemas as expressing means–end links: if an applicable schema links from an action to a state that is more positively valued than the states linked to by the negation of that action, the schema influences the choice machine to take that action. A further provision allows schemas to *compose* together, such that if C and D are now true, the schemas

$$C{:}A{\rightarrow}B\ (x),\quad C{:}{\sim}A{\rightarrow}B\ (y),\quad D{:}B{\rightarrow}R\ (v),\quad D{:}{\sim}B{\rightarrow}R\ (w)\ (x > y,\ v > w)$$

assuming they are not currently overridden, combine to imply that action A is a means to achieving B, and that the action of achieving B in turn achieves R.[18] Hence, given CD now, A's utility with respect to R now is $u(x\text{-}y)(v\text{-}w)$ (making the usual conditional-independence presumptions).

5.6.2 The Evidentialist Problem with Schemas

The machinery postulated so far is still within the evidentialist paradigm—the machinery effectively presumes that a kind of evidential link is a

17. The expected utility must also be scaled by the probability of the schema's applicability. In the examples here, I make the simplifying idealization that it is certain that a schema's context conditions are satisfied (or are not) at each given moment.

18. Composing schemas together (forming the composite actions mentioned above in sec. 2.4.1) allows the choice machine to anticipate the outcome of an action in a perhaps previously unencountered situation (e.g., *CD* in the above example), even though each schema individually tabulates statistics over actually encountered situations. Other such combinatorial techniques (e.g., as proposed in *Made-Up Minds*) are also useful, but are not needed for the present analysis.

means–end link. As such, the machinery stipulated so far is vulnerable to the street-crossing problem discussed in the previous section, given the idealizations proposed there.

To demonstrate that vulnerability, say the choice machine has the schemas

$*SD{:}C \rightarrow \sim T_2$ (safely disposed crossing)

$*SD{:}\sim C \rightarrow T_2$ (0.3)

with the various predicates (Street, Cross, Disposition, Traffic) defined as in section 5.5. The schemas assert that if the (safely disposed) choice machine crosses the street, no dangerous traffic then arrives; if it does not cross, dangerous traffic may arrive. The asterisks are to denote that—intuitively, and by the criteria outlined below—the schemas are misleading if construed as expressing a means–end link from crossing, to the traffic condition, rather than just a correlation between the two. That is, although the conditional probabilities expressed by these schemas are correct as conditional probabilities—which address what is in fact the case when the specified conditions are in fact met—it is mistaken to presume that these are also the probabilities that there *would* be dangerous traffic if a safely disposed agent *were* to cross now, or if it were not to cross (in the choice-supporting sense of *would*). Although C reliably gives evidence of $\sim T_2$ (given SD), C is not a means to achieving $\sim T_2$. But the choice machinery, by virtue of the utility calculations proposed so far, does presume a means–end link whenever schemas assert an evidential link. The machinery needs a way to reject that presumption here.

Suppose S, D, and T_1 are true of the current situation. Thus, the above schemas' contexts are satisfied in the current situation, and (let us suppose) the choice machine has found extensive, exceptionless empirical support for these schemas. How could the choice machine know that the schemas nevertheless do not express a means–end link from action to result? That is, how could it know that crossing the street now—with oncoming dangerous traffic—would not achieve the absence of dangerous traffic at the moment of crossing?

Suppose further that the choice machine has the schemas

$ST_1{:}C \rightarrow T_2$ (conserved traffic)

$S{\sim}T_1{:}C \rightarrow \sim T_2$ (conserved nontraffic)

which assert that the dangerous traffic neither vanishes nor materializes when the agent crosses. The conserved-traffic schema, in particular, contradicts the prediction made by the safely-disposed-crossing schema when both schemas are applicable (i.e., when SDT_1 is true). But the conserved-traffic schema does not meet the criteria for exception-overriding the safely-disposed-crossing schema; it is not more specifically conditioned. Just as in section 5.5 (where we considered the evidential conflict between how my safely disposed street-crossing always in fact proceeds—safely and without dangerous traffic—and how oncoming dangerous traffic always in fact proceeds—without suddenly vanishing), the choice machine needs some additional principle to resolve the conflict, to determine that the conserved-traffic schema is the one to trust—even though both conflicting schemas enjoy exceptionless empirical confirmation. (Empirical support for $ST_1{:}C{\rightarrow}T_2$ could be found in situations where agents of unsafe disposition—i.e., D is false—do cross in front of dangerous traffic.)

We would like the machinery to be able to exception-override the safely-disposed-crossing schema $^{*}SD{:}C{\rightarrow}{\sim}T_2$ with one that says

$^{**}SDT_1{:}C{\rightarrow}T_2$.

That is, although the action of (initiating) crossing the street (with safe disposition) ordinarily implies the absence of dangerous traffic passing at that moment (T_2), if crossing occurs under this strictly more-specific condition—that is, in the presence of dangerous oncoming traffic (T_1)—then the traffic T_2 is still present. Unfortunately, the probability expressed by this schema is undefined (as denoted by the double asterisk), because SDT_1C has zero probability, as previously discussed. For the same reason, the choice machine can obtain no empirical support for this schema (C never occurs when SDT_1). Indeed, as discussed in section 5.5, that empirical support would be lacking even in the case of a nonzero but sufficiently minuscule probability of SDT_1C. And there is mutually contradictory empirical evidence that relies on different subsets of the conditions SDT_1C—namely, SDC (given which T_2 never actually occurs) and ST_1 (given which T_2 always occurs). These considerations impugn any unexplained subjective probability conditioned on SDT_1C.

5.6.3 The Explaining-Away Principle: Restraining Evidentialism

Despite the foregoing problems, there is a plausible basis for trusting the conserved-traffic schema $ST_1{:}C{\rightarrow}T_2$ over the safely-disposed-crossing

schema *$SD{:}C{\rightarrow}{\sim}T_2$. Intuitively, we'd like to be able to say that in situations where the safely-disposed-crossing schema is applicable and crossing occurs, the only reason dangerous traffic doesn't arrive—just as the safely-disposed-crossing schema said it would not—is because (in those situations) no dangerous traffic approached to begin with, as the conserved-*non*traffic schema $S{\sim}T_1{:}C{\rightarrow}{\sim}T_2$ asserts. Therefore, when that reason doesn't obtain—namely, when dangerous traffic does approach, so the conserved-nontraffic schema is inapplicable—there's no reason to believe what the safely-disposed-crossing schema says in that situation. The *explaining-away* principle tries to formalize that intuition, as follows.

A key observation is that the conserved-nontraffic schema $S{\sim}T_1{:}C{\rightarrow}{\sim}T_2$ is both *more general than* and *explanatory of* the safely-disposed-crossing schema *$SD{:}C{\rightarrow}{\sim}T_2$, in the following sense:

- Define a given schema to be *more general than* another if they share the same action conditions, and if the given schema has been activated (i.e., its action conditions have obtained when its context conditions obtained) in a strictly wider set of circumstances than those in which the other schema has been activated. Here, the conserved-nontraffic schema is more general than the safely-disposed-crossing schema because the latter is applicable only when D (hence activated only when D). And although the former schema is applicable only when ${\sim}T_1$, nonetheless the latter schema too has never been (indeed cannot be) activated except when ${\sim}T_1$ (i.e., SDT_1C is an impossibility under the proposed idealizations).
- Define a given schema to be *explanatory of* another if it predicts the same result of the same action with (approximately) the same reliability.

Intuitively, when a given schema is explained by a more general schema in the foregoing sense, an Occam's-razor presumption suggests that the given schema is just a consequence of the explanatory schema, owing its apparent validity to the explanatory schema, and so should not be insisted on as expressing an independent principle.[19] The explained schema should just defer to the explanatory schema, and so should be considered inapplicable (even though its context is satisfied and no exception-override obtains), letting the explanatory schema do its work instead. Call this provision an

19. Baum (2004) argues compellingly that Occam's razor, in various guises, is central to our cognitive machinery's acquisition and representation of knowledge.

explaining away of the explained schema. I propose that the explaining-away principle be built into the choice machinery, revising the earlier definition of a schema's applicability.[20]

Suppose the explained schema has its context conditions satisfied, and the explanatory schema is applicable. The explanatory schema (by definition) reiterates the explained schema's prediction, so the explained schema's deferral has no noticeable effect. But if instead the explanatory schema is *not* currently applicable, then the explained schema's deferral matters (assuming there are no other applicable schemas that also assert the same prediction as the explained schema). For in that case, if there is another applicable schema asserting a conflicting prediction, that other prediction becomes uncontested because its competitor was explained away. So the predictive ambiguity resolves in favor of that other schema.

Thus, for example, as a consequence of being explained away, the schema $*SD{:}C{\rightarrow}{\sim}T_2$ does not contribute (as it otherwise would) to the calculation of C's utility with regard to ${\sim}T_2$ when S and D are true. Instead, that explained schema defers to the explanatory schema $S{\sim}T_1{:}C{\rightarrow}{\sim}T_2$, which makes the same prediction. But when T_1 is true, that explanatory schema is inapplicable, and so the conflicting applicable schema $ST_1{:}C{\rightarrow}T_2$ prevails instead, with no applicable schema contradicting it. Thus, even though the explained schema $*SD{:}C{\rightarrow}{\sim}T_2$ is exceptionless (given our zero-probability idealization), and is empirically recognized as such by the choice machinery, the machinery does not construe that schema as expressing a means–end link.

The explaining-away principle tugs in a different direction than the exception-override, giving priority to more general schemas rather than more specific ones. Their rationales are complementary:

- When a schema has in fact been activated in a specific circumstance (enough times to obtain a significant sample), an expectation of the then-observed outcome takes precedence (in future occurrences of that specific circumstance) over a conflicting prediction about what is expected to occur more generally; hence, the exception-override.

20. A more advanced version of this principle would allow explanation jointly by multiple more-general schemas. For example, $A_1{:}B_1{\rightarrow}C_1$ and $A_2{:}B_2{\rightarrow}C_2$ are each more general than $A_1A_2{:}B_1B_2{\rightarrow}C_1C_2$, and they jointly explain the latter schema (presuming conditional independence). But the present simpler principle suffices for the examples here.

▪ The explaining-away principle, in contrast, addresses schemas that are in *agreement* about what is expected to occur. But the deferral shifts the context for that expectation to that of the more-general explanatory schema. Given a specific circumstance in which an explained schema has *not* been activated (e.g., no agent has stepped in front of clearly seen dangerous oncoming traffic while having the contrary desire and competence)—so there is no empirical basis for an exception-override regarding that specific circumstance—the deferral of an otherwise-applicable schema to a currently inapplicable explanatory schema allows any conflicting applicable schema to prevail uncontestedly, thus effectively adjudicating between two reliable (even exceptionless) conflicting schemas.

5.6.4 Would-ness

As already noted, the safely-disposed-crossing schema $*SD{:}C{\rightarrow}{\sim}T_2$ is impeccable as an expression of conditional probability. On the idealizing assumptions above, 100 percent of actual *SDC* situations also exhibit ${\sim}T_2$. Given *D*, the action *C*—the action of (initiating) crossing the street—is indeed perfect evidence for the absence of dangerous passing traffic. No adjustment or override of the conditional probability per se is warranted, nor of the expected utility defined by the product of that conditional probability and any utility ascribed to the schema's result condition ${\sim}T_2$.

Nonetheless, it would be nonsensical, given T_1, to use that expected utility to assess the desirability of choosing action *C*—that is, to treat the evidential relation between *C* and ${\sim}T_2$ (although perfectly valid as such) as a means–end relation.[21] The explaining-away principle suppresses that means–end construal, countering the preliminary presumption that a schema's evidential link also corresponds to a means–end link, thus addressing the evidentialist dilemma raised in section 5.5.

21. Allais (Allais and Hagen 1979) calls attention to other ways that people's decision intuitions diverge from what the maximization of expected utility would dictate. But that divergence involves situations with substantial uncertainty; Allais shows that people often place a premium on the predictability of a desired outcome, even at the cost of a lower expected utility. In contrast, the present distinction contradicts expected-utility-based decisions even when uncertainty is negligible (or 0). The present decision approach still selects an action based on the product of a conditional probability and a utility value (as does an expected-utility calculation), but the present approach substitutes a subjunctive probability for a conventional conditional probability in that calculation.

The explaining-away principle thus enables the choice machine to distinguish, in some situations, between the conditional probability

$$\Pr(Result|Context\&Action) = \Pr(Result\&Context\&Action)/\Pr(Context\&Action)$$

—the probability that the result conditions are actually the case given that the context and action conditions actually are—and the *subjunctive* (or *modal* or *counterfactual*) probability

$$\Pr(Result\backslash Action|Context),$$

where $\Pr(A\backslash B|C)$ is the probability that A *would* be the case now if B were (in the choice-supporting sense of *would*), given that C actually is the case. (Similarly, e.g., Pearl 2000 distinguishes $\Pr(A|BC)$ from the subjunctive $\Pr(A|do(B),C)$, but Pearl's version only refers to what B would *cause*; see sec. 5.6.3 below. Similarly too with Gibbard and Harper 1977.) Because of the explaining-away principle, the machinery computes the utility of a contemplated action with respect to the subjunctive probabilities of various outcomes, rather than with respect to their (conventionally defined) conditional probabilities.

The machinery proposed here does not compute subjunctive probabilities as such. Rather, each schema keeps track only of an associated conditional probability. Whenever the schema's context is satisfied and the schema is not exception overridden, the associated conditional probability is presumed also to be the subjunctive probability, *except* when an explaining away defeats that presumption and the schema defers to a more general explanatory schema. At such times, the subjunctive probability is obtained instead from other, nondeferred schemas, if available.[22] Thus, in lieu of proposing a mathematical formula for the value of a subjunctive probability,

22. There may be situations where no schema is applicable, or where multiple such schemas remain in mutual contradiction not resolved by the explaining-away principle (or by further machinery proposed below in sec. 6.2.1). The choice machinery needs to be designed with heuristics to let it muddle through such situations. But those heuristics address situations in which the choice machine's understanding is manifestly inadequate or confused (or, conceivably, in which there is genuine indeterminacy as to a particular counterfactual consequent). So we should not expect the heuristics to render a decisive judgment about what's true about the counterfactual consequents in those situations. Not so, however, with the conflicting predictions as to what would happen if the action of crossing were taken in the street-crossing scenario. That conflict does have an unambiguous resolution that the machinery should be able to find.

I am sketching a presumption-and-deferral mechanism for computing (some) subjunctive-probability values.

The central challenge of a theory of subjunctive reasoning is to find a principled way to determine what actually true propositions to ''hold fixed,'' and what propositions to ''let vary'' instead, for consistency with a counterfactual antecedent. For example, when contemplating the hypothetical antecedent C given DT_1, do we ask what must be true assuming D would be the same as it actually is, or assuming instead that T_1 would be the same as it actually is? The present proposal reduces that challenge to a problem of adjudicating among conflicting inductive projections (e.g., those given by the safely-disposed-crossing schema and by the conserved-traffic schema) in the absence of a direct empirical resolution (both schemas are never activated simultaneously), a problem addressed by the explaining-away principle.[23] (Sec. 5.6.5 briefly discusses a contrasting approach to the problem of subjunctive ambiguity, Lewis's *possible-worlds* analysis.)

Clearly, the explaining-away principle applies as well to a wide class of everyday problems with the same structure as the street-crossing problem—including problems that would have been ubiquitous during our ancestors' evolution. I speculate that some such principle built into our choice machinery is what makes it intuitively obvious to us that my crossing the street, although perfectly evidential of the absence of dangerous traffic, *would* not yield that condition right now (as the traffic approaches), and thus does not now serve as a means to that end.

If the explaining-away principle were not built in, we might still reason (as above) that the principle is a good one to use. That explicit reasoning would conflict with, but not necessarily prevail over, the influence of, say, the misleading street-crossing schema that tells us it is safe to cross no matter what (assuming that our choice machinery is generally schemalike in the first place). The explicit reasoning would not necessarily prevail because

23. Suppose there can be rare (but not yet encountered) occasions when both schemas *are* activated simultaneously. Explaining away then helps resolve conflicts not only in subjunctive reasoning (what would be the case if the action were now taken?) but also in inductive projection (what *is* the case when the action is in fact taken?). (In *Made-Up Minds*, chap. 8, I propose that a *deductive override*—an earlier formulation of the present explaining-away principle—might help resolve the induction riddle illustrated by the famous *grue* paradox of Goodman [1983].)

(at least according to the present account) the explicit reasoning is itself implemented by some complicated network of schemas whose very terms of representation designate abstractions belonging to the theory (abstractions such as correlation, explaining away, and so forth). Even if those schemas weigh in on the question of the subjunctive consequence under consideration here, they're just another instance of schemas making a competing claim about that consequence; no decisive resolution of the conflict would be available (see n. 22 above). Even though those schemas may express a theory that says how to resolve the competing claims in question, the schemas would not somehow alter the underlying machinery to ensure that it functions accordingly (just as, say, an acquired physical characteristic does not somehow retrofit itself into an organism's genes).

Plausibly, causal regularities correspond to the most widely applicable schemas we can find, because we inhabit a universe where a small number of such regularities combine in exponentially many configurations to specify all that occurs. The lower the level of abstraction, the more widely applicable are the regularities. For example, principles about the behavior of gears and levers apply more widely than principles about the black-box behavior of a particular machine made up of some specific arrangement of gears and levers, because the components individually occur far more widely than in some particular combination—and similarly for components consisting of molecules, atoms, quarks, and so forth.

The explaining-away principle favors widely applicable explanatory schemas; those are the schemas that explain away others, rather than vice versa. Insofar as causal schemas are especially widely applicable and explanatory, schemas that correspond to causal links would resist being explained away—except by schemas that express lower-level causal principles. The explaining-away principle thus promotes a reductionist presumption that the (more widely applicable) regularities of the constituent parts of an object determine the expectation of the object's behavior in circumstances where that behavior has not been tested directly.

The foregoing is not a knockdown argument that explaining away favors recognizing causal links as means–end links. It is just a defeasible plausibility argument to that effect. But if that argument is right, then it is not only causal regularities that the explaining-away principle allows to be recognized as means–end links. Define the predicates

U (Up) I take the action of lifting up my hand now.

R (Raised) My hand is raised.

P_R (Past, Raised) The state of the universe one billion years ago is such that, according to correct physical laws, my hand is raised one billion years thence.

and suppose that the choice machinery includes the (empty-context, hence unconditional) schemas

$:U \rightarrow R$

$:\sim U \rightarrow \sim R$

$:U \rightarrow P_R$

$:\sim U \rightarrow \sim P_R$.

Since R is true exactly when P_R is true (presuming determinism), empirical support for the first pair of schemas will be identical to support for the second pair. No circumstances will provide an exception override or an explaining away of the second pair without doing so for the first pair. The choice machinery will recognize a means–end link to P_R to the same extent as to R (and appropriately so, as argued in sec. 5.4) even though the former link is acausal.

The recognition of causal relationships as such is quite plausibly a more sophisticated task than the recognition of subjunctive relations generally, if the latter recognition can be performed by something like the machinery sketched here. In that case, there is no reason to expect that our built-in choice machinery is (or should be) designed to treat only causal links as means–end links. An exclusion of acausal means–end links would be hard to implement (if specifically causal relations are hard to recognize as such)[24] and would be of no benefit. Indeed, the exclusion would only *impair* an agent's ability to pursue some of its goals—both whimsical goals like the hand-raising past predicate and (as argued below) the more impor-

24. There are, however, specific aspects of causal links whose recognition is plausibly hardwired. For example, well-known experiments show that young infants are (at some level) aware that an object's response to another object's motion requires physical contact at the time of the response (e.g., Flavell and Markman 1983). But such perceptually based special-case criteria are distinct from a more general basis for ascertaining causal relations among phenomena that may be novel, abstract, or widely dispersed in space and time.

tant goals that arise in Newcomb's Problem (esoterically) and in Prisoner's Dilemma situations (routinely and crucially).

I propose, then, that a choice-supporting subjunctive link just *is* (something like) an evidential link that does not get explained away, in the technical sense defined here.[25] That is what (I claim) would-ness, in the choice-supporting sense, turns out to consist of; that is what a means–end relation turns out to consist of. As the street-crossing example illustrates, an agent that was designed (or that evolved) to use that subjunctive criterion of means–end relations would thereby fare well in achieving its goals, compared to an agent that used a purely evidential criterion—thus passing the metacircular-consistency test with regard to that comparison. And as the hand-raising example illustrates, going to the extra trouble of distinguishing specifically causal relations, and insisting on their use alone as means–end links, would not improve an agent's goal pursuit, and indeed would sometimes hamper it, compared to using the subjunctive criterion. So the subjunctive criterion passes the metacircular-consistency test with regard to that comparison as well.

5.6.5 Contrasts: Lewis's Possible Worlds, and Pearl's Causality

A competing approach to analyzing subjunctive or counterfactual inference appeals to *possible worlds* (e.g., Lewis 1973). By that approach, what would be the case if I were to cross the street now (as dangerous traffic plainly approaches, and in fact I do not cross) is whatever *is* the case in imaginary alternative possible worlds in which I do cross the street now, but which, given that difference from the actual world, are otherwise as similar as possible to the actual world.

But everyday intuitive measures of similarity give wrong answers (see Fine 1975). For instance, a possible world in which there is a momentary lull in traffic right now (so that I now cross) would thereby differ from the

25. The parenthetical qualification refers to several hedges. First, section 6.2 below proposes an additional principle that adjusts what gets construed as a means–end link. Second, as mentioned above in note 20, a choice machine would benefit from a more advanced version of the explaining-away principle—a version whereby several more-general schemas can combine to be explanatory of another schema. Third (a catchall), the present proposal is preliminary and tentative. Even if it proves to be a step in the right direction, refinements are no doubt needed.

actual world in a much more ordinary way than would a possible world in which I cross now, despite the danger and despite my safe disposition, because some cosmic rays (or whatever) induce a bizarre sudden disruption of my street-crossing competence. Construing the more-ordinary difference as smaller leads to the conclusion that if I were to cross now, the dangerous traffic would be absent. But that's obviously wrong, at least with regard to the choice-supporting sense of *would*—the sense in which we rationally act for the sake of what would then be the case.

Alternative similarity criteria proposed by Lewis (1979b) and others hinge in some way on the physical extent of the differences just prior to the action. Altering the current traffic is a physically bulkier change than tweaking a few of my neurons to make me cross despite the danger; in that sense, the former difference from the actual world is indeed larger than the latter, which supports the correct conclusion. But the relative physical bulk of the two changes is an inessential feature of the scenario; it is easy to contrive different scenarios of the same structure but where the traffic equivalent happens to be physically smaller than the choice machinery, leading to the opposite, incorrect conclusion about what would happen.

The possible-worlds approach to subjunctive inference is not to be confused with the so-called multiple worlds of Everett's formulation of quantum mechanics (chap. 4 above)—although Deutsch (1997) proposes that (physically real) alternative quantum branches do in fact serve as the possible worlds supposedly referred to by counterfactual assertions. But Deutsch gives no reason for that construal, other than to assert that the specification of counterfactual consequents would be arbitrary without that grounding in physical reality. On the contrary, though, there are many conceivable ways to ground that specification in (some abstraction of) physical reality—including the approach advocated here, which outlines how our choice machinery might compute subjunctive inferences from data gleaned from our ordinary experiences, rather than requiring access to arcane aspects of physics. Moreover, Deutsch's approach too is tripped up by the sort of counterexample just mentioned: in a different scenario where the traffic equivalent happens to be much smaller than the choice machinery, a quantum branch in which the traffic equivalent undergoes a bizarre disruption may be less unlikely (i.e., may receive more

quantum weight) than a branch in which the choice machinery suffers such a disruption.

Judea Pearl's *Causality* (2000) offers an alternative, purely causal theory of how to derive decisions from probabilities (expressed in Bayes nets) and utilities. (The explaining-away principle presented here can perhaps be seen as a simplification of Pearl's preference for so-called minimal latent structures.)

Because some requisite conditional probabilities are undefined, Pearl's derivation of causal models from Bayes nets does not apply to scenarios in which full determinism is presumed, such as the hand-raising example above. If, however, we allow a minuscule probability that past-state P_R does not lead to hand-raising action U or that U does not lead to raised-hand R, then (even if such deviations are so overwhelmingly improbable that they never occur in the lifetime of the universe) Pearl's approach is applicable, and it can show correctly that P_R causes U and U causes R, but U does not cause P_R. But then, an agent using only causal links as means–end links (as Pearl advocates) would needlessly forfeit the opportunity to have P_R be true if P_R is a goal (with similar squandered opportunities in Newcomb's Problem and the Prisoner's Dilemma, as discussed below in chapters 6 and 7).

Pearl (ibid., p. 108) distinguishes an *act* (''a consequence of an agent's beliefs, disposition, and environmental inputs'') from an *action* (''an option of choice in contemplated decision making, usually involving comparison of consequences''). Pearl models the latter as uncaused: if represented as a node in a causal structure, an action has no parent nodes (p. 71). An action is a ''free choice'' (p. 109) or a ''surgical intervention'' (an externally originated change to the causal structure itself) with respect to which Pearl calculates the subjunctive probability of an outcome (rather than using the conditional probability of an outcome given an act) to assess a contemplated action's utility.

But the crucial distinction between choice-supporting subjunctive probability and evidentialist conditional probability does not require Pearl's further distinction between an act and an uncaused action. In fact, a choice is both an act and a *caused* action (as Pearl defines the two terms), and can be so modeled for decision-making purposes. The difference is both

philosophical—in effect, Pearl's formalism models free will rather than mechanical choice—and practical, in that Pearl fails to recognize what I argue are (in some situations) valid acausal means–end links.

By modeling actions as uncaused, Pearl places the actor beyond the scope of the system's mechanical rules, much as the Copenhagen interpretation of quantum mechanics so places the observer (recall chap. 4 above). In both cases, I argue, this special exemption for ourselves (as observers or as actors) is unwarranted and distorts our understanding of our role in the universe.

5.7 Summary

In a deterministic universe, no outcome ever changes (from what it was already set to be). All is inalterable, sitting statically in spacetime. Still, contemplating the operation of a choice machine (sec. 5.3) in a deterministic world, we see how an agent can sensibly act on means–end links in such a world, taking actions for the sake of goals.

Even though an action cannot change an outcome from whatever it is predetermined to be, we can draw a contrast between actual situations where the action occurs, and other actual situations where it does not. Correlations found in such contrasts support an evidentialist presumption of a means–end link.

Sometimes, as in the past-predicate example of section 5.4, a means–end link is acausal, but still correct (though there is not, of course, some sort of acausal ''force''; the link is just an abstract relation). An agent that fails to recognize and act on such a means–end link forgoes the achievement of its goal, a needless fatalist resignation.

But often, as in the street-crossing example of section 5.5, a mere correlation—even an exceptionless one—is *not* a means–end link. An agent misconstruing it as such would not achieve its goal either. Inalterability does not imply futility, nor does acausality. But conversely, mere evidence does not ensure efficacy, and this chapter's explaining-away principle says why, providing a way to selectively defeat the evidentialist presumption. Arguably, incorporating the explaining-away principle into an agent's means–end machinery transforms evidential means–end recognition into subjunctive means–end recognition—the agent selects an action on the basis of what would then be the case, in the choice-supporting sense of

would, as elaborated above. This subjunctive criterion is broader than a causal criterion but narrower than mere correlation.

The means–end recognizing machinery proposed so far turns out to have yet another basic technical problem, not addressed by the explaining-away principle and requiring an additional principle to fix it. This further problem directly reflects the intuition that it is futile to act for the sake of what your action cannot change—the intuition that underpins the perceived incompatibility of choice and determinism. The next chapter's discussion (sec. 6.2) of a variation of Newcomb's Problem—with both boxes transparent—highlights the additional technical difficulty and proposes a technical solution. First, though, to set up for that discussion, the next chapter begins by analyzing the conventional, opaque-box version of Newcomb's Problem, using the means–end machinery as described so far.

6 Deterministic Choice, Part 2: Newcomb's Problem and Beyond

6.1 Newcomb's Problem

The preceding chapter argued that an agent rationally acts for the sake of what would then be the case, in the choice-supporting sense of *would*. A means–end link is thus a subjunctive link, which, as just argued, is intermediate between an evidential link and a causal link. This chapter returns to Newcomb's Problem, introduced above in section 5.1: you are given an opaque box that either contains $1,000,000 or is empty; and you are offered an additional, transparent box containing a visible $1,000. Earlier, a reliable prediction ascertained whether you would accept or decline the transparent-box offer; if and only if you were predicted to decline, $1,000,000 was placed in the (now-sealed) opaque box. You are informed of these circumstances before you make your choice regarding the transparent box.

In this section, I apply to Newcomb's Problem the notion of a subjunctive, possibly acausal means–end link, in an attempt to resolve the paradox created by two conflicting, intuitively compelling prescriptions:

- Take just the opaque box because that box would then be found to contain $1,000,000 more than otherwise; or,
- Take both boxes because you then obtain $1,000 more than otherwise, in addition to however much money the opaque box already inalterably contains.

I argue in this chapter that the paradox revealed in Newcomb's Problem is at the root of the seeming conflict between determinism and meaningful choice in general. Resolving the paradox in Newcomb's Problem points the way to understanding how choice and determinism are compatible after all.

Suppose we elaborate Newcomb's Problem with the following science-fiction assumptions. A while ago, a mischievous benefactor, preparing to present the two boxes to you, took a detailed snapshot of the nearby state of the universe, and then used that snapshot to run a (perhaps atom-by-atom) simulation of the subsequently unfolding events, up to and including your forthcoming choice. If the simulation showed you choosing both boxes, the benefactor put nothing in the opaque box; if it showed you choosing the opaque box alone, the benefactor put $1,000,000 in that box. As usual, let us make the idealization that the universe is deterministic enough, and predictable enough even in practice, to carry out the simulation with perfect reliability (or, recast the scenario in an artificial world where those idealizations hold). Let us also consider an alternative situation in which the simulation is fallible, such that, say,

$$\Pr(M|N{\sim}P_B) = \Pr({\sim}M|NP_B) = x < 1,$$

defining the indexical assertions

N (Newcomb) An agent is now presented with a Newcomb's Problem choice.

B (Both) The agent chooses both boxes now.

P_B (Past, Both) The past state of the universe at the time of the snapshot is such that (according to correct physical laws) the agent will take both boxes when presented with the forthcoming choice.

M (Million) The opaque box presented to the agent contains $1,000,000.

For the sake of argument, assume money has linear utility (e.g., having twice as much is twice as good) and assume that the expected monetary payoff is the only relevant goal here.

One way to respond to the conflict between the two prescriptions in Newcomb's Problem is to propose that the scenario is logically impossible, that there is an inherent contradiction in the very assumption that such a scenario could be realized. In that case, the argument for taking just the opaque box, and the argument for taking both boxes, could both follow correctly from the premises describing the scenario. There would then be no paradox—Newcomb's Problem would merely show, unsurprisingly, that contradictory premises lead to contradictory conclusions. I begin this section by arguing that the Newcomb's Problem scenario is *not* logically

impossible, and that the apparent paradox therefore does need to be resolved.

The opaque box's opacity serves a critical technical function. If the box were transparent, its then-visible content could affect the outcome of your choice process. The visible content would also, in effect, announce a prediction of your choice process. As in the discussion in section 5.3, the simulation would therefore need to simulate the effect of the box content on your choice process. But the simulation does not yet know what the box content will be (the content depends on the still-pending outcome of the simulation itself), so the simulation would need to simulate itself at that point (to find out what the box content will be). But that self-simulation will in turn reach the point where it needs to know the simulation outcome in order to proceed further, requiring a self-self-simulation, and so on—a crippling infinite regress.

However, because the box is opaque—and assuming more generally that until you make your choice, you are sufficiently insulated from any effect caused by the content of the box or by the process or outcome of the simulation—it is possible to conduct the simulation in such a way as to leave the content of the box unspecified, and still be able to simulate what goes on outside the box, and in particular what goes on in your choice process.

There may seem to be a risk of infinite regress despite the box's opacity, however. After all, your deliberation might well include trying to anticipate what is in the box by anticipating what the simulation predicted. You would, in effect, try to simulate the simulator (not an atom-by-atom simulation, of course, but rather at a much higher level of abstraction). In so doing, you would then be simulating the simulator's simulation of your simulation, and so on. Your simulation thus could never reach its conclusion—and neither, it seems, would the benefactor's simulation, as it simulates your own nonterminating simulation.

But let us assume that the rules of the encounter impose a time limit, even if a generous one—you're allowed, say, up to a year to puzzle it out and make your choice; if you don't explicitly choose by then, you're construed by default to have taken both boxes. We've stipulated that the benefactor had conducted a perhaps atom-by-atom simulation of you and your surroundings, simulating much faster than the actual events unfold. Your own simulation of the simulation, therefore, would have to be much

slower. Regardless of the semantic content of your deliberation, the benefactor's simulator needs only to track the course of a finite number of particles for a finite time (i.e., the particles that implement your deliberation and that implement the surrounding environment) in order to anticipate the conclusion of your deliberation. Therefore, the simulation encounters no infinite regress. Thus, your engaging in a circular simulation (until you run out of time) would be an inadvisable tactic (what good would it do you?), but in any case it would be no impediment to the benefactor's accurate simulation.

But even if there is thus no infinite-regress problem, *chaos theory* (see Gleick 1987) shows that long-range prediction is sometimes prohibitively impractical, even in some deterministic systems, because arbitrarily small differences in initial conditions can eventually have arbitrarily large effects on what unfolds (such systems are called *chaotic*), and you can't know the initial conditions *exactly*. Nonetheless, a great deal that unfolds in the world is predictable in practice with reasonable reliability. (Sometimes even human choice in Newcomb's Problem and Prisoner's Dilemma situations can be reliably predicted using informal folk psychology, as discussed in sec. 7.1; and complex artificial environments—potentially including agents making humanlike choices—are entirely practical to simulate, as discussed just below.) Chaos or quantum mechanics may sometimes make prediction impractical or physically impossible, but for purposes of this thought experiment we are making the (logically possible) assumption that the particles can be accurately simulated.

Qualitatively, the Newcomb's Problem prediction resembles weather forecasting, where an array of meteorological measurements and geographic data constitute an approximate snapshot of the local state of the universe. That snapshot is input to a program that reliably (at least for a while) simulates events faster than they actually unfold. There, too, the terrestrial system being forecast encompasses the forecasting simulation itself, raising the possibility of an infinite regress. For instance, large-scale human activity may have immediate effects on weather patterns; different levels of traffic and industrial activity may cause weekday weather to differ on average from that of the weekend, for example. Conceivably, this large-scale activity could be influenced by the forecast itself, in turn influencing the weather. But if we sufficiently insulate the predicted system from the prediction itself—for example, by not publicly divulging the forecast until

the prediction expires—then that influence can be rendered negligible, so the simulation can ignore it without loss of accuracy. Similarly, in Newcomb's Problem, the box's opacity helps insulate the prediction (as revealed by the box content) from the predicted system until after the predicted event occurs. (In practice, of course, a weather prediction's influence on the predicted system is likely negligible even without any effort to conceal the prediction; this is not so in Newcomb's Problem.)

One way to show the logical possibility of the requisite simulation and insulation is to recast Newcomb's Problem in terms of a particular target computer running a program that implements a rich environment replete with choice-making agents, one of which is about to be presented with a Newcomb's Problem choice. Suppose I take a snapshot of the digital state of the target computer and input that state to a simulator running on another, much faster computer. Suppose that from the moment of the snapshot until after the chooser has chosen, the target computer program accepts no inputs except for a single bit transmitted just before the boxes are set up. This bit controls whether the opaque box (which is part of the environment implemented by the program) will be set up to be empty or to contain $1,000,000; if and only if the simulator predicts that the chooser will take just the opaque box, the transmitted bit specifies that $1,000,000 will be put in that box.

Until the opaque box is subsequently opened, the target computer program does not allow the box's content to affect any part of the environment external to the box (although arbitrarily complex effects may occur *inside* the box; for example, a recording device within the box might monitor the box content to document that the content did not change after the initial setup). The simulation can thus predict what choice the chooser will make, by simulating the entire digital environment implemented by the target computer program, except for the (yet-unknown) opaque-box interior. Because the box's exterior is insulated from any effect of the interior, the simulation can ignore the interior with no loss of fidelity.

We can easily enact this scenario today using ordinary computer technology (though not yet with choice-making agents as intelligent as people, of course). Neither the simulation nor the insulation is perfect:

- There is always a minuscule chance that a hardware error will disrupt the execution of the target computer's program, in which case the simulation of the target computer may be inaccurate.

▪ The physical process of the simulation, and of the target computer's representation of the opaque-box content, cannot have literally zero effect on the rest of the target computer's circuitry, including whichever neighboring atoms in that circuitry help represent the choice-making agent. But in a properly functioning computer, any such effect is negligible, and disappears altogether at the digital level of abstraction, the level at which we regard the circuitry as running a computer program.

Thus, the prediction and insulation are *almost* perfect. The chance of an erroneous prediction, or of leakage through the insulation, can be made arbitrarily small.

Returning now to the original scenario, with a human chooser instead of a digital agent, the presumed atom-by-atom simulation and insulation present much greater technical difficulties, far beyond current technology and perhaps even forever insurmountable. Still, the difference is just one of degree. There is no logical impossibility to performing the requisite simulation with arbitrary accuracy, even in the case of a human chooser. And logical possibility is all we need here. Practicality is beside the point, because the purpose of contemplating Newcomb's Problem is not to prepare for the eventuality of actually finding ourselves in such a situation. Rather, the purpose is to explore some of the ramifications of choice given determinism—in particular, the ramification that sometimes, we can sensibly act for the sake of an already inalterable outcome even if we have no causal influence over that outcome. Chapter 7 below argues that much the same reasoning turns out to apply to practical, real-world Prisoner's Dilemma situations.

The logical possibility of a Newcomb's Problem scenario obliges us to resolve the paradox between the arguments for and against taking just the opaque box; the paradox does not just stem from contradictory premises. Making the case for taking only the opaque box in Newcomb's Problem—even with a fallible simulation—is relatively straightforward; the challenge is to invalidate the arguments *against* the one-box choice. This challenge reflects the general caution about paradoxes discussed in section 1.2.3 above. When arguments conflict, creating a paradox, at least one of the arguments must be wrong. But we cannot defeat the incorrect argument merely by countering it with a sound argument for the opposite

conclusion—that's what constitutes a paradox in the first place. We must also show what is wrong with the incorrect argument, by identifying a step in the argument that does not follow.

The argument for taking just the one box is that if you were to do so, you would (probably or certainly, depending on how reliable a simulator we postulate) obtain $1,000,000 in the opaque box; if you were to take both boxes, you would (probably or certainly) obtain nothing in the opaque box. Accordingly, using the same calculation that we would use if your choice *caused* the opaque-box content to *change*, we find that someone who would take both boxes has a much lower expected gain from the encounter than someone who would take just the opaque box—even if the simulation's reliability is only, say, 0.9, or even 0.6 or less.

The intuitive appeal of taking just the opaque box can be boosted dramatically—especially for a highly reliable simulation—if we imagine that you first engage in a long series of practice trials, using play money. During the practice trials, you vary your choice whimsically just to see what the outcome will be. Say you then find a million (play) dollars in the opaque box whenever you take that box alone; when you take both boxes, you always find the opaque box empty. After such a demonstration, and in the absence of any reason to doubt that the real-money trial works by the same rules, are you seriously going to choose both boxes when the stakes are real?

One-box choosers do get a better payoff than two-box choosers (or, in the probabilistic case, they get a better payoff on average). Still, this observation is not decisive. A skeptic might maintain (as do Gibbard and Harper [1977], for example) that taking both boxes is the rational choice, and if one-box choosers fare better than two-box choosers, it is only because the situation has been rigged in advance to reward the former's (predicted) irrationality (much as if a written exam were perversely graded to reward mistaken answers).

But this skeptical position is suspect unless (unlike in the present situation) the chooser is unaware of the rigging, and thus unable to take it into account when choosing (as the exam taker could do if she knew the grading scheme). Otherwise, a committed fatalist in a deterministic universe could maintain that the entirety of spacetime is effectively a grand Newcomb's box whose content—including the outcome of every choice—is

already inalterably sealed in. Those who insist (irrationally, says the fatalist) on making choices in pursuit of goals do fare much better, on average, than those who succumb to fatalist resignation. But (concludes the fatalist) that contrast just shows that the universe's rules are rigged in advance—though not necessarily by any deliberate design—in a manner that rewards the irrational.

As discussed in section 5.2, means–end criteria cannot be deduced without some built-in starting point, so the fatalist cannot be definitively refuted. Still, if we put aside complete fatalist skepticism, acknowledging instead that there is meaningful choice, even given determinism (as sec. 5.3's choice-machine discussion argues), then the door is opened to a means–end link to an already-inalterable result. And section 5.4 presented an acausal link—from the action of raising my hand, to the satisfaction of the hand-raising past-predicate—that serves as a means–end link to a past state, to the same extent that there is also a means–end link to the (also already-inalterable, given determinism) state of my hand being elevated a moment from now.

Hence, neither a choice's or a result's determinism or inalterability, nor the acausality of a link from a choice to a result, is necessarily disqualifying. And indeed, Newcomb's problem harnesses a past-predicate link (to the snapshot-time state of the universe) quite similar to the hand-raising past predicate, followed by an ordinary causal link from the past state to the box's content (via the benefactor's simulation, etc.).

A more difficult challenge to the one-box choice stems from the question of what a means–end link does consist of, if it is not just a causal link. A popular version of the one-box argument merely appeals to what would *have to* be true of the opaque-box content (or, for a fallible predictor, what would *likely* be true) if the one-box choice were made, compared to if the two-box choice were made, given the way the opaque-box content was set up. But similarly, in the street-crossing scenario, the absence of dangerous traffic would *have to* be the case if I were to cross now, given my safe disposition. As discussed in sections 5.5 and 5.6, that criterion is merely evidential, and can be met even in the absence of a means–end link. Thus, a version of the one-box argument that appeals to an evidential link is invalidated by the correct objection that the existence of a means–end link does not follow just from the existence of an evidential link.

The previous section's discussion of the street-crossing scenario—and of the subjunctive means–end criterion implemented by schemas with explaining-away deferrals—shows how to circumvent that objection. I propose a different version of the one-box argument; I maintain that the acausal link in Newcomb's Problem is a means–end link by virtue of more than just being an evidential link. To elaborate, let us define

N_{100} The Newcomb simulator is 100 percent reliable.

N_{99} The Newcomb simulator is 99 percent reliable (i.e., if B, there is a 0.99 chance that the simulator predicts the both-boxes choice; and if $\sim B$, there is a 0.99 chance that the simulator predicts the one-box choice).

in addition to the predicates N (Newcomb), B (Both), P_B (Past, Both), and M (Million) introduced above. And suppose the choice machinery has the following schemas:

$N{:}B{\rightarrow}P_B$ $\quad$ $N{:}{\sim}B{\rightarrow}{\sim}P_B$

$NN_{100}{:}P_B{\rightarrow}{\sim}M$ $\quad$ $NN_{100}{:}{\sim}P_B{\rightarrow}M$

$NN_{99}{:}P_B{\rightarrow}{\sim}M$ (0.99) $\quad$ $NN_{99}{:}{\sim}P_B{\rightarrow}M$ (0.99).

The first two schemas constitute a past-predicate link similar to the hand-raising example: if and only if you were to take just the opaque box, the snapshot-time past state would have been such that (according to physics) you will so choose. The second or third pair of schemas (whichever pair is applicable, supposing that either N_{100} or N_{99} is true) corresponds to an ordinary causal link from the past-predicate state to the box content: the past state causes the snapshot to cause the simulation to cause there to be $1,000,000 in the opaque box. As argued in section 5.6.4, causal links are especially general (the more so the lower level they are), hence resistant to being explained away; and similarly for links to result conditions that are past-predicates corresponding to causal result conditions.

In the absence of an explaining away of the above schemas, they compose together (recall the composition provision in sec. 5.6.1) to establish, by the proposed subjunctive criterion, a means–end link from the one-box choice, to having $1,000,000 in the opaque box: if you were to choose just one box, there would (more likely) be $1,000,000 in the opaque box, in the choice-supporting sense of *would*, despite the lack of a causal link from your choice to the box content. The link is not merely evidential. By the proposed criterion, the means–end link, although acausal, does not exist

merely by virtue of a correlation, but also by virtue of the absence of a (currently inapplicable) more-general explanatory schema for the above past-predicate link or for the ordinary causal link. That is, the means–end link exists by virtue of the absence of schemas that could explain away one or both of those constituent links.

Instead of appealing to the two pairs of composed schemas above, it would be more straightforward just to invoke the complementary schemas $N{:}B{\rightarrow}{\sim}M$ and $N{:}{\sim}B{\rightarrow}M$, which directly express the link from the box choice to the opaque-box content. But these schemas express a correlation that, even if exceptionless, is neither causal, nor a link to a past-predicate that corresponds to a causal result. Hence, we have no reason to be confident that these more directly formulated schemas are resistant to being explained away.[1] By itself, then, the merely evidential correlation expressed by the directly formulated Newcomb's schemas does not assure a means–end link. Instead, we need to reconstruct that correlation by composing the above past-predicate link and causal link, each of which, plausibly, is indeed resistant to being explained away.

Newcomb's Problem and the street-crossing scenario have similar structure in terms of both the causal and the probabilistic dependencies among their respective states. In each scenario, there is a common causal influence (past-predicate ${\sim}P_B$ in Newcomb's Problem, or traffic-calmness ${\sim}T_1$ in the street-crossing scenario) on both a choice (taking just one box, or crossing the street) and a goal state ($1,000,000, or traffic nonarrival ${\sim}T_2$), creating a correlation between choice and goal, but with no causal link from choice to goal. What distinguishes the scenarios, by the present account, is the comparative explanatory generality (in the sense defined) of the conflicting schemas in each scenario. The explaining-away principle, appealing to that distinction, permits an acausal evidential link to stand as a means–

1. Indeed, if we *either* presume that $1,000,000 was put in the opaque box, or presume that no money was put there, then the directly formulated Newcomb's schemas can be explained away by conserved-box-content schemas, much as the safely-disposed-crossing schema was explained away by a conserved-nontraffic schema in section 5.6.3. A plausible extension of the choice-machinery discussed here would support a case-by-case analysis of an unknown condition (such as the opaque-box content). In the current problem, then, the directly formulated Newcomb's schemas would be explained away in each of the two possible cases.

end link in Newcomb's Problem, while (fortunately) preventing a similar construal in the street-crossing scenario.

The antievidentialist challenge to the one-box stance is answered by the argument above that the means–end link here is not merely evidential; rather, it is a subjunctive link, which fulfills additional criteria. Another, even deeper challenge arises from a suggestion by Nozick (1969) to boost the *dominance* argument for the two-box choice—the argument that the box already contains either $1,000,000 or nothing, and either way, you do better (by $1,000) if you take both boxes than if not.[2]

Suppose a video hookup permits a friend of yours to see inside the opaque box while you contemplate your choice. Your friend is remotely located and has no way to communicate with you until after you have chosen. Your friend has been briefed about the scenario's rules. Looking at the video image, your friend either sees $1,000,000, or sees nothing, in the opaque box. Either way, your friend concludes that your more lucrative choice is to take both boxes, thereby gaining $1,000 in addition to the visible $1,000,000 (or in addition to the visible $0, as the case may be).

Your own expectation about the box content is contingent on your expectation about your choice, but your friend's perception of the box content is unconditional. Your friend not only has strictly more knowledge than you of the situation—she also has all the knowledge of the situation that is needed to answer the question at hand: which is the more lucrative choice for you to make? (I.e., which is the choice for which the payoff would be greater than if the other choice were made?) She might realize that, from the perspective of your more limited knowledge, the other choice might reasonably seem more lucrative. But she, from her perspective, knows (if she is reasoning correctly) which choice *is* the more lucrative one.

2. Nozick (1969) endorsed the one-box choice only in the case of an infallible predictor, on the grounds that the dominance argument does not apply if only one possible choice is consistent with the (albeit yet-unknown-to-you) box content. But with even an infinitesimal chance of simulator error, Nozick saw no way to counter the dominance argument, though he acknowledged his discomfort with that dependence on the difference between a zero probability and an arbitrarily small one. Later, Nozick (1993) revised his position to advocate using a weighted sum of the apparently intractably conflicting decision strategies.

But here a problem arises regarding the correctness of the one-box choice—even its correctness from your own (limited) perspective. The problem is that you yourself can know—if you recapitulate the above reasoning—that your friend, if she is reasoning correctly, must have concluded that taking both boxes would be your more lucrative choice. You need not know what she sees in the box to know that (if she is reasoning correctly) she has reached that conclusion—because you know she would reach it *regardless* of what she sees in the box. But if you know that she has all the information needed to identify the more lucrative choice, and you know what her conclusion must be if she is reasoning correctly from that information, then you thereby know what the correct conclusion is in fact—that is, that taking both boxes is more lucrative than taking just the opaque box.

Note that even if you are persuaded by this argument for the two-box choice, you still face the contradictory argument above for the one-box choice—the argument that there is a means–end link from taking just the opaque box, to the corresponding past-universe state; and from there, to $1,000,000 in the box. (And recall the play-money practice-trials scenario to bolster the intuitive appeal of the one-box choice; in the case of a highly reliable simulator, you know from direct experience that you will find $1,000,000 in the box if and only if you take that box alone.) The peeking-friend argument *counters* that means–end analysis, but does not *invalidate* it (recall the distinction discussed in sec. 1.2.3)—that is, the peeking-friend argument does not identify any step in the means–end analysis that does not follow properly from the previous step, and thus does not resolve the paradoxical conflict between the two arguments. I continue to defend the one-box choice; in the next section, I argue that it is instead the peeking-friend argument that turns out to be flawed.

There is a way to escape the force of the peeking-friend argument. The one-box choice can be correct only if, contrary to the foregoing, your friend somehow would *not* be reasoning correctly, given what she sees in the box, to conclude that taking both boxes would actually be your more lucrative choice. But in that case, we may as well let the opaque box be transparent, along with the $1,000 box. If (somehow) your friend would be correct to conclude that the one-box choice would be more lucrative for you, despite her seeing what is in the box, then you yourself would

(somehow) be correct to reach the same conclusion, even if the box content were visible to *you* as well.

I accept this reduction of the opaque-box version of the problem to the transparent-boxes version. (Indeed, if, as I maintain, the one-box choice is correct in the opaque-box version of the problem, then—as Gibbard and Harper [1977] point out—in principle you can understand the problem well enough to be certain that you will so choose, and—in the case of a reliable simulator—you can thus know in advance that the box contains $1,000,000, even though the box is opaque.[3]) I argue in the next section that—as counterintuitive as it seems at first—the one-box choice remains correct in the transparent-boxes case. I argue that the dominance argument is invalid here, that inalterability does not imply futility even when the already-inalterable state is also already visible, and that the mistaken contrary intuition leads as well to the perceived incompatibility of choice and determinism.

3. Along these lines, Eells (1982) and Jeffrey (1983) provide an evidentialist argument for taking both boxes in Newcomb's Problem (the opaque-box version, with a reliable predictor)—in contrast with the usual, more straightforward evidentialist argument for the one-box choice—by appeal to screening off via knowledge of one's decision just prior to acting (recall n. 12, chap. 5). First, they assume that the agent's final decision to take just one box might (rarely) result inadvertently in the action of taking both, or vice versa (thus averting the zero-denominator problem that would otherwise prevent the requisite conditional probabilities from being defined). Then, they argue, given a decision to take both boxes, the opaque box is already known (almost certainly) to be empty; and if it is empty, it must remain so whether or not the action does accord with that decision. Or, given a decision to take just the large box, the opaque box is known (almost certainly) to hold $1,000,000, and again does not change regardless of the action itself. Thus, knowledge of either decision (just prior to the action itself) screens off knowledge of the box content from knowledge of the action, just as though the box were transparent—given knowledge of the decision (and hence knowledge of the box content), the action itself provides no (further) evidence about the box content, so there is no evidentialist link between the action and the content.

But as argued in section 5.5 and chapter 5, note 12, evidentialism (even with screening off by knowledge of the decision prior to acting) gives the wrong answer even in some mundane situations. (Screening off by foreknowledge of the decision also leads to the fatalist street-crossing problem discussed below in sec. 6.2.1.) If evidentialism is thus unfounded, its prescription in Newcomb's Problem is moot, so it doesn't matter whether or not Jaffe and Eells are right about what that prescription is.

6.2 Newcomb's Problem with Transparent Boxes

The argument so far is that fatalist (sec. 5.3) and causal (sec. 5.4) means–end criteria are too strict; contrary to those criteria, inalterability does not imply futility. But an evidentialist criterion (sec. 5.5) is too lax, sometimes mistaking mere correlations for means–end links. Section 5.6 describes instead a subjunctive criterion. The proposed machinery for recognizing subjunctive means–end links makes an initial evidentialist presumption, but uses an explaining-away principle to sometimes override the initial presumption.

The analysis of Newcomb's Problem (sec. 6.1) appeals to the proposed subjunctive means–end criterion, but the explaining-away principle does not intervene there. Rather, the argument is that the key evidential links in Newcomb's Problem (namely, a causal link and a particular kind of past-predicate link) are the sort of links that the explaining-away principle does *not* override, so those presumed means–end links just remain standing. Still, the explaining-away principle is important to the analysis, because the proposal to presume evidential links to be means–end links in general—in Newcomb's Problem and elsewhere—would be untenable, due to the wrong answers it gives in many mundane situations (e.g., the street-crossing scenario of sec. 5.5), if the explaining-away principle were not available to correct the presumption when the presumption clearly errs.

Nozick's peeking-friend scenario challenges the foregoing analysis of Newcomb's Problem, arguing that the correct choice is no different from what it would be if the already-fixed box content were also already known; thus, the argument concludes, the one-box choice is wrong (a conclusion that, as noted above, leaves a paradox, an unresolved conflict with the still-standing means–end analysis that argues for the one-box choice). To explore the implications of Nozick's scenario, I now consider a variant of Newcomb's Problem in which *both* boxes are transparent, so that the content in question *is* already known to the chooser. I argue that the one-box choice is still correct, and that this variant requires us to confront head-on the intuition that says we need not bother to act for the sake of what is already inalterable—in particular, to act for the sake of the now-transparent box's content; but also, in a deterministic universe, to act for the sake of *anything*.

Making both boxes transparent in Newcomb's Problem poses an immediate technical difficulty—the infinite regress discussed in section 6.1. If a prediction of your choice will be communicated to you before you make the choice (in this case, communicated via the now-visible box content), then the prediction cannot necessarily be made in the first place. For in the course of the simulation, the simulator reaches the point where the prediction is conveyed to you. The simulator is then unable to proceed because the prediction itself is still in progress, its conclusion not yet known.

But this technical problem admits of a technical solution. The simulation can tentatively *presume* that the prediction will be that you take only one box. Accordingly, the simulation shows $1,000,000 in a (now also transparent) large box. The simulation proceeds, and if it then shows you taking just the $1,000,000 large box, then the one-box prediction is made, and the $1,000,000 large box is presented to you in reality (along with $1,000 in a smaller transparent box). The situation then is just as the simulation had tentatively presumed it would be, so the prediction predicated on that presumption will be accurate.[4]

But if the simulation instead shows you taking both boxes (when confronted with $1,000,000 in the large box), then an empty large transparent box is presented to you in reality (along with $1,000 in a small transparent box). Thus, in this version of the problem, we give up on requiring that an *empty* large box, if presented to you, is predictive of the choice you then make, because the simulation does not consider the empty-box case (but see sec. 6.3 below for another variation of the problem). So if you find yourself presented with an empty large box, you have no good reason to refrain from taking both boxes. (We assume, of course, that you are accurately informed of the revised rules for this revised version of the problem.)

Consider, first, the case of a perfectly or almost-perfectly reliable simulation. Suppose you were to find yourself presented with $1,000,000 in the large box, reflecting a prediction that you will take just the large box. Should you then choose just the large box, or both boxes?

The large box's visible content in effect already tells you what you are about to do, which may seem to imply that no choice remains for you

4. We continue to stipulate, as in the opaque-box version of the problem, that you are insulated from any influence caused by the simulation process—except, of course, for the now-visible large-box content, which effectively announces the simulation outcome to you.

to make. On the contrary, though, as in the deterministic street-crossing scenario with oncoming traffic—where your choice is also a foregone conclusion—the mechanical choice process itself continues to operate, and crucially so: you compare what would be the case if you took one action or another, and you act according to which state of affairs you prefer. The box content may effectively inform you in advance of your choice (as does the oncoming traffic in the street-crossing scenario), but the box content does not somehow suppress your choice process to impose the announced choice on you (and neither does the visible oncoming traffic, though it is certainly an influential *input* to the choice process). Rather, you still just choose whichever action you think is best. So you must still figure out which choice *is* the best.

But how then could the one-box prediction reflected by the visible $1,000,000 possibly be correct (given the stipulation that maximizing your profit is the sole motivation for your forthcoming choice)? What could conceivably stop you from taking the extra $1,000 (and thus rendering the one-box prediction false)?

If you believe (as most people would) that taking both boxes would be your more lucrative choice, then indeed nothing would stop you from doing so—which is why you would not in fact find yourself presented with $1,000,000 in the first place in such an encounter; rather, the large box would be empty. But if you were convinced (at least by the time you make your choice) that taking just the large box is the correct, more lucrative choice—which, I claim, is the case, counterintuitively enough—then *that conviction* is what would lead you to take just the large box. It is not a matter of somehow being restrained from making what you believe is the better choice; rather, it is a matter of believing (correctly, I claim) that taking just the large box *is* the better choice. But it remains for me now to justify that very surprising claim.

I argue that there is the same reason to forgo the extra $1,000 here as there is in the earlier, opaque-box scenario: if and only if you were to take both boxes, the simulation would (almost certainly) have so predicted; and if and only if the simulation were to so predict, there would be no money in the large box (even though, in fact, there *is* $1,000,000 in the box, which cannot now change). These two subjunctive links compose together to link your choice to the box content. The rest of this section explores this argument in more detail.

One-box choosers do fare better in transparent-boxes encounters, for it is they who are presented with $1,000,000 in the first place. Still, the recommendation that it is best here to take just the large box, even though that box is transparent and you already see $1,000,000 inside, is so violently counterintuitive as to make the original formulation of the problem seem almost banal.[5] Evidentialists and causalists alike would of course take both boxes here:

- For causalists, the large-box content is irrelevant to your choice, since (as in the opaque-box version of the problem) your choice has no causal influence on that content. But taking both boxes would cause you to receive an extra $1,000.
- For evidentialists, the expected payoff, given that you take both boxes, is $1,001,000, compared to a mere $1,000,000 given the other choice (conditioning in both cases on the visible, hence already known, large-box content).

(Thus, by the way, I am proposing here a subjunctive link—between your choice and the large-box content—where there is not even an evidential link, not even a correlation. This proposal is contrary to my earlier claim that the subjunctive criterion is broader than a causal criterion but stricter than an evidential criterion. A more careful formulation of that earlier claim is that the subjunctive means–end links given by individual pairs of complementary schemas—one schema with a given action, the other with its negation, both schemas with the same context and result—are indeed intermediate between correlation and causation; but a composition of two or more such links, as in the present example, can form a subjunctive link that is not even an evidential link.)

Anecdotally, I find that even many individuals who agree with the one-box choice in the original version of Newcomb's Problem—who are therefore persuaded that it can sometimes make sense to act for the sake of an already determined outcome, even in the absence of a causal link—still balk at the one-box choice in the transparent-boxes scenario. Let us examine what makes the one-box recommendation seem so bizarre here, and why (I claim) the intuitions it counters are mistaken.

5. Gibbard and Harper (1977) mention a transparent-boxes variation (though with no discussion of a simulation-based prediction, or the infinite-regress problem and its solution), and advocate taking both boxes. Kavka's *toxin problem* is also relevant; see section 6.2.4 below.

Just as the explaining-away principle is proposed in order to repair what would otherwise be an unwarranted laxness in the proposed means–end-recognizing machinery, this section introduces another principle, the *prejudiced-context* principle, to keep the machinery from being too strict, to keep it from succumbing to fatalism by failing to recognize means–end links to outcomes that are foreknown. I digress again to the street-crossing scenario to motivate this new principle by appeal to a class of mundane situations in a deterministic universe. I then return to the transparent-boxes variant of Newcomb's Problem to apply the new principle to that problem.

6.2.1 Foreknowledge in the Street-Crossing Scenario

Foreknowledge of an already determined choice, and of its outcome, presents a difficulty beyond the problem addressed in section 5.6, the problem of correlations that are not means–end links. That previous problem was addressed by appeal to the explaining-away principle. The new difficulty is especially conspicuous in the transparent-boxes variation of Newcomb's Problem, but it appears as well in mundane, real-life situations like street-crossing, presuming determinism. The difficulty, I believe, goes to the heart of the fatalist intuition that choice and determinism are incompatible. The difficulty arises as follows.

Returning to the street-crossing scenario, recall the predicates S (Street), C (Cross), D (Disposition), T_1 (Traffic, imminently arriving), and T_2 (Traffic, arrived). Suppose the choice machine also uses the predicates

O (Other) The agent reaches the other side of the street shortly.

H (Hit) The agent is hit by traffic shortly.

H_1 (Hit_1) The state of the world now (including the agent) is such that (according to correct physical laws) the agent is not hit by traffic shortly.

and suppose the choice machine has the schemas

$S{\sim}T_1{:}C{\rightarrow}O{\sim}H$ (successful crossing)

$S{:}{\sim}C{\rightarrow}{\sim}O{\sim}H$ (noncrossing)

$^{*}S{\sim}H_1{:}C{\rightarrow}O{\sim}H$ (fatalist)

plus the four empty-context schemas

$:H_1{\rightarrow}H,\quad :H{\rightarrow}H_1,\quad :{\sim}H_1{\rightarrow}{\sim}H,\quad :{\sim}H{\rightarrow}{\sim}H_1.$

Each of the above schemas expresses an exceptionless relation between its action and result, given its context. The fatalist schema, however, turns out to be misleading (hence the asterisk) if dangerous traffic is just about to arrive (i.e., if T_1 is true). The fatalist schema asserts that if I cross the street in a situation where I am not in fact about to get hit by traffic, then I reach the other side—and am not hit by traffic. I thus achieve both my primary goal (not being hit) and my secondary goal (getting to the other side). The combined utility of the two result conditions of this currently applicable schema exceeds the utility of the action $\sim C$ (which achieves just the primary goal), so C is seemingly the preferable action, given $S{\sim}H_1$.

And in fact, we *are* given $S{\sim}H_1$. That is, under the operative idealizations, I already know—even with the dangerous traffic oncoming now—that the event of my getting hit by traffic in the next moment is not in fact present in the spacetime of the actual universe. I already know that simply because I already know that I will not cross now in the presence of such traffic. And the actual absence of that event of getting hit (regardless of the *reason* for its absence) is all that $\sim H$ or $\sim H_1$ asserts. So the context $S{\sim}H_1$ is indeed satisfied, and I know it. Thus, if I were to (mis)construe this schema's relation between action and result as a means–end connection now, a gravely incorrect action would follow.

Moreover, let us assume that there is a minuscule physical possibility of crossing successfully and without collision even in the presence of dangerous traffic (the several lanes of densely packed speeding cars *might* all manage to miss me). If I cross the street now, in front of the onrushing traffic, the result condition of the (actually exceptionless) fatalist schema $*S{\sim}H_1{:}C{\rightarrow}O{\sim}H$ could then follow without logical contradiction—it is (barely) possible that I'd indeed escape collision. But of course it would still be a grave error to count on that result and to act accordingly, since the result would be very unlikely—contrary to the fatalist schema's misleading assertion.

Since the fatalist schema $*S{\sim}H_1{:}C{\rightarrow}O{\sim}H$ holds without exception, there is no prospect of an exception override for that schema. In particular, there is no additional context condition under which crossing when not in fact about to be hit results in being hit; no situation exhibiting an exception to the result $\sim H$ will ever be encountered when the context condition ($\sim H_1$) is satisfied.

Moreover, unlike the misleading schema *$SD{:}C{\rightarrow}{\sim}T_2$ in section 5.5, the fatalist schema *$S{\sim}H_1{:}C{\rightarrow}O{\sim}H$ is not explained by a more-general schema that is currently inapplicable. Hence, the fatalist schema does not get explained away. In particular, the fatalist schema does not defer to an explanation by the successful-crossing schema $S{\sim}T_1{:}C{\rightarrow}O{\sim}H$, because $S{\sim}T_1C$ is not more general than $S{\sim}H_1C$. Indeed, *$S{\sim}H_1{:}C{\rightarrow}O{\sim}H$ corresponds to a causal link—the action of crossing, when (by sheer luck) it does occur under the specified circumstances, does indeed cause reaching the other side (without collision). As discussed in section 5.6.4, causal links tend to be maximally general and explanatory, so we would not expect that schema to be explained away and thus to defer to any other schemas that contradict its prediction.

Even if we suppose that the choice machine has the schema

$ST_1{:}C{\rightarrow}H{\sim}O$ (0.99) (dangerous crossing)

which correctly predicts the outcome of crossing in front of oncoming dangerous traffic—namely, a strong likelihood of being hit and not reaching the other side—this dangerous-crossing schema neither exception-overrides nor explains away the fatalist schema, so the choice machinery (as specified so far) does not know which of those conflicting schemas to trust. The dangerous-crossing schema does not create an overriding exception, because the schema's activation does not constitute a special case of the fatalist schema's circumstances of activation (on the contrary, the two schemas' activations are almost mutually exclusive). It does not constitute a more-general explanatory schema, because it asserts a *contrary* result to that of the fatalist schema, and thus certainly does not explain any occurrence of the fatalist schema's result.

The schema *$S{\sim}H_1{:}C{\rightarrow}O{\sim}H$, if misconstrued as a means–end link when T_1, crystallizes the oft-perceived incompatibility between choice and determinism. The schema captures the fatalist intuition that asks rhetorically: since the past already determines that I'm not actually about to get hit by traffic now, why bother doing anything (such as not crossing) for the sake of that already guaranteed goal? Since (it is already determined that) ${\sim}H$ is true—and since I already know it—I can (seemingly, according to the misleading schema) cross now safely and successfully, despite the dangerous oncoming traffic. The foreseeable predetermination of the outcome ${\sim}H$ appears to lead to the (obviously absurd) conclusion that crossing now

would be safe—that is, that it would be futile for me to act (by not crossing) for the sake of the already inalterably achieved goal of not being hit by traffic.[6] That apparent conclusion from the foreseeable predetermination is seemingly a reductio ad absurdum of the foreseeable predetermination.

The considerations put forth in section 5.3 above offer a rebuttal to this fatalist alleged ramification of determinism; we can see that a choice machine described in section 5.3 could operate usefully in street-crossing scenarios and the like, even in an artificial world that is clearly deterministic. Hence, we should not conclude fatalistically that determinism renders choice futile. The point now, though, is that the particular analytical choice machinery proposed here, as specified so far, is itself vulnerable to drawing a fatalist conclusion from foreseeable predetermination, and so is in need of repair.

There is an obvious intuitive reply to this seeming incompatibility between choice and determinism. Yes, I am not in fact about to be hit by traffic. Yes, the present state of the universe already inalterably assures that; yes, I already know that the present state already assures that. And yes, anytime I actually cross when I am not in fact about to be hit by traffic, I reach the other side—and (tautologically) am not hit. Nonetheless, I *would* (contrary to actual, already known fact) very likely be hit (and the present state would, contrary to already known fact, be such that I will very likely be hit) if, also contrary to already known fact, I were to cross now (in front of the onrushing dangerous traffic).

This intuitive reply translates into a solution for a choice machine. An additional principle needs to be built into the machinery. To repeat the problem as the proposed machinery stands so far: the fatalist schema $*S{\sim}H_1{:}C{\rightarrow}O{\sim}H$ is applicable when $S{\sim}H_1$ is true. That schema is not exception-overridden or vulnerable to being explained away, even when T_1 is true. The schema thus motivates taking action C for the sake of $O{\sim}H$ (since, as the above noncrossing schema correctly asserts, the alternative, $\sim C$, leads to $\sim O{\sim}H$, a result of lesser utility, presuming the goal of getting

6. Of course, a further application of the same reasoning would also argue for the futility of crossing for the sake of the (also already determined, one way or the other) outcome of getting to the other side. But fighting fatalism with fatalism will not produce a reasonable decision strategy.

to the other side). To block the inappropriate influence of the fatalist schema when T_1 is true, I propose to augment the mechanism with a built-in *prejudiced-context* principle, as follows.

Suppose the fatalist schema $*S{\sim}H_1{:}C{\rightarrow}O{\sim}H$ is now applicable. The problem with this schema is that when T_1 is true, the truth of the context condition ${\sim}H_1$ implicitly depends on the fact that C does not now occur. The condition ${\sim}H_1$ is indeed true, of course; otherwise, the schema would not now be applicable. But if a schema's context thus effectively prejudges whether the schema's action even occurs, the asserted action-contingent result may be distorted. The prejudiced-context principle is designed to identify and override such a schema, in the following manner.

Suppose there are other applicable schemas that would be activated now if a given schema were activated, and according to which the given schema wouldn't have been applicable now in the first place if its action were taken now. Here, the applicable schema $ST_1{:}C{\rightarrow}H{\sim}O$ (0.99), which composes with $:H{\rightarrow}H_1$, would be activated if the given schema $S{\sim}H_1{:}C{\rightarrow}O{\sim}H$ were activated. $ST_1{:}C{\rightarrow}H{\sim}O$ is activated if C now occurs, predicting the result H; and in turn, $:H{\rightarrow}H_1$ is activated if H occurs, predicting the result H_1. Thus, if the given schema's action C occurs, these other applicable schemas culminate in probable result condition H_1 that contradicts the given schema's context $S{\sim}H_1$, thus contradicting that schema's very applicablity. (To clarify, we have a contradiction only because both occurrences of H_1 refer to the same event. In contrast, when schemas designate a temporal change—e.g., $X_{t1}{:}A{\rightarrow}{\sim}X_{t2}$—there is no contradiction between context and result.)

Define a prejudiced-context schema as a given schema for which other schemas (1) would be activated if the given schema were, and (2) predict results that contradict the context of the given schema. I propose that a prejudiced-context schema be treated as currently inapplicable. It thus does not contribute to its designated action's current utility. In particular, then, the schema $*S{\sim}H_1{:}C{\rightarrow}O{\sim}H$ is a prejudiced-context schema, and as such is treated as inapplicable, solving the problem of its fatalist influence.

Two aspects of the prejudiced-context principle's rationale bear elaboration:

- Although the motivating intuition is that a prejudiced-context schema's context prejudges whether the action occurs, the proposed recognition of a prejudiced-context schema runs in the other direction: the schema's action (with the help of other schemas) subjunctively predicts a contradic-

tion of the context. To see why a prediction of the action by the context wouldn't be a correct criterion instead, consider the schema $ST_1{:}C{\rightarrow}H{\sim}O$ (0.99); it is not (and should not be) considered a prejudiced-context schema according to the proposed rule even if safe disposition D is true and even if D and ST_1 (with the help of other schemas, such as perhaps $:SDT_1{\rightarrow}{\sim}C$) predict that the action C does not occur when the schema is applicable.

▪ If not for the prejudiced-context principle, the same fatalism problem would arise—and perhaps more straightforwardly—if the result condition H itself were substituted for precursor condition H_1 in the context of the schema $*S{\sim}H_1{:}C{\rightarrow}O{\sim}H$. In that case, though, an obvious proposal would be to rule out by fiat either the inclusion of a designated result condition in a context, or else the inclusion of a context condition that is not temporally subsequent to the result condition(s). Using precursor condition H_1 in the context is meant to demonstrate that a temporal-ordering or context-result-disjointness rule would be easily circumvented and therefore unhelpful; and a rule imposing a temporal ordering would sometimes be harmful, in that it would block the recognition of what section 5.4 argued are legitimate means–end links to some past states. (A rule barring from the context any condition strongly *correlated* with a designated result condition would also be undesirable; for example, such a rule would exclude even ${\sim}T_1$ from the context of $S{\sim}T_1{:}C{\rightarrow}O{\sim}H$, because the absence of onrushing dangerous traffic correlates strongly with reaching the other side of the street.)

The prejudiced-context and explaining-away principles are complementary; neither suffices in the absence of the other:

▪ As already discussed, explaining away (and exception-override) fail to disable the fatalist schema $*S{\sim}H_1{:}C{\rightarrow}O{\sim}H$. We also need the prejudiced-context principle.

▪ Conversely, if not for explaining away, the prejudiced-context principle would be vacuously ambiguous, because we need other schemas to reveal the prejudiced schema's context's implicit presumption about the schema's own action; and (if not for explaining away) the prejudiced-context schema could make those other schemas seem as though *they* were instead prejudiced-context schemas, as in the following example.

As discussed in section 5.6.3, explaining away rescues the choice machine from the safely-disposed-crossing schema $*SD{:}C{\rightarrow}{\sim}T_2$. Without

the explaining away, *SD:C→∼T_2, composed with :∼T_2→∼T_1 (assuming the existence of that exceptionless schema too), asserts that T_1 wouldn't hold if C were to occur now. So according to those schemas, the context of ST_1:C→H∼O is contradicted if that schema's action occurs now. Thus, without explaining away, *SD:C→∼T_2 would incorrectly make ST_1:C→H∼O seem to be a prejudiced-context schema—thus invalidating the latter schema and thereby preventing it from (correctly) tagging *S∼H_1:C→O∼H as a prejudiced-context schema. Either schema could thus be invalidated, depending on which one was first deemed to be a prejudiced-context schema.

The explaining-away principle resolves that ambiguity by invalidating *SD:C→∼T_2 on other grounds, preserving the schema ST_1:C→H∼O, and thus preserving that schema's effect on the status of *S∼H_1:C→O∼H. (To make this resolution work properly, we need to stipulate that the explaining-away principle is applied *before* the prejudiced-context principle is applied.)[7]

The explaining-away principle deals with a schema that has not been (or cannot be) tested with respect to a given overriding condition—for example, *SD:C→∼T_2 cannot be tested (by activating it) when T_1 obtains (because disposition D prevents crossing when T_1). The prejudiced-context principle deals with a schema whose context incorporates the negation of a condition that is in fact a consequence—in the subjunctive sense, not necessarily causal—of the schema's activation, even if there is no possible overriding condition under which the schema's result fails to obtain (e.g., with the fatalist schema *S∼H_1:C→O∼H, there can be no additional condition under which crossing when not about to be hit results in being hit).

Explaining away is intended to keep the means–end-recognizing criterion from being too lax, and thus protect the choice machine from a wildly exaggerated sense of the efficacy of its actions. The prejudiced-context prin-

7. Pinning down the sense in which the decision calculation should not condition on a state that "depends on" an action is tied to the long-standing problem of statistical *confounding*. Proposed resolutions easily fall prey to the circularity just exhibited: you don't know which "dependent" conditions to rule out in the case of a particular conditional probability until you have already answered the same question for other conditional probabilities—but that answer in turn first requires answering the original question. (See, e.g., Pearl 2000, chap. 6, for his proposed resolution in his causal decision theory.)

ciple is intended to keep the criterion from being too strict (in that, e.g., the fatalist schema's bogus means–end link from C to $O{\sim}H$, holding $\sim H$ constant, is effectively a denial of the actual means–end link from C to H or from $\sim C$ to $\sim H$). The prejudiced-context principle is thus meant to protect the choice machine from fatalist complacency about an already-known outcome that the machine still in fact has a choice about.

6.2.2 Foreknowledge in Newcomb's Problem

The analysis of Newcomb's Problem with transparent boxes is exactly parallel to the foregoing. We confront the fatalist intuition that the visible $1,000,000 (like the world-state that guarantees I am not just about to be hit by traffic) is already known to be inalterably there; and given its presence, taking both boxes seemingly results in achieving both the primary (already inalterably secured) goal of getting the $1,000,000 in the large box, and the secondary goal of getting the $1,000 in the small box (just as crossing the street—given that in fact I am not about to be hit—seemingly results in the primary, already assured goal of not getting hit, plus the secondary goal of getting to the other side). In contrast, taking just the large box results in only the primary goal of getting $1,000,000, forfeiting the $1,000 (just as not crossing still meets the primary goal of safety, but forfeits the secondary goal of reaching the other side). So it is apparently futile to act by refraining from taking the small box (or by refraining from crossing) for the sake of the already inalterable, already known to be secured primary goal, at the cost of forfeiting the secondary goal.

But that appearance is false; the action is not futile. If (contrary to fact) the alternative action were taken, then (also contrary to fact) the (albeit already known to be inalterably secured) primary goal would *not* be achieved. Of course, this subjunctive link does not mean that the already secured goal—in Newcomb's Problem or the street-crossing problem—ever *changes* its state, in any actual situation, from one moment to the next, when the action is taken. On the contrary, the presence (or absence) of $1,000,000 in the box does not change at the moment of choice. Likewise, there is never any change, at the moment of choice, in the already past universe state (nor, therefore, is there any change in that past state's assurance about whether I will imminently be hit by traffic).

Just as in the opaque-box case, the means–end link in the transparent-boxes problem comes from the composition of two other such links:

- If (contrary to fact) you were to take both boxes, then (also contrary to fact) the snapshot-time past state of the universe would be such that (according to physics) you will choose both boxes if presented with $1,000,000 in the large box. Just as in the hand-raising past-predicate example, you do not cause the past to be what it was. Yet you do have a choice about this particular aspect of the past, exactly as you have a choice about which action to take now.
- If (contrary to fact) that past state were such that you will choose both boxes if presented with $1,000,000 in the large box, then (also contrary to fact) the benefactor would put no money in the large box. This latter means–end link is conventionally causal, mediated by the simulation.

Similarly, if you were to take just the large box, the past state would be such that (according to physics) you will take only that box; and if the past state were thus, there would be $1,000,000 in the box (as in fact there is). Achieving that aspect of the past state is a means to the goal of there being $1,000,000 in the large box.

Schemas expressing those two composed-together means–end links (similar to the schemas in the previous section) describe the structure here just as in the opaque-box version of the problem:

$NM_1{:}B{\rightarrow}P_B$, $NM_1{:}{\sim}B{\rightarrow}{\sim}P_B$,

$NN_{100}{:}P_B{\rightarrow}{\sim}M_1$, $NN_{100}{:}{\sim}P_B{\rightarrow}M_1$,

$NN_{99}{:}P_B{\rightarrow}{\sim}M_1$ (0.99), $NN_{99}{:}{\sim}P_B{\rightarrow}M_1$ (0.99),

$:M_1{\rightarrow}M$, $:{\sim}M_1{\rightarrow}{\sim}M$,

where we define some of the predicates slightly differently than in section 6.1:

N An agent is now presented with a transparent-boxes Newcomb's Problem choice.

P_B The past state of the universe at the time of the snapshot is such that (according to physics) if the agent sees $1,000,000 in the large box, the agent will take both boxes when presented with the forthcoming choice.

M_1 The large box presented to the agent contains $1,000,000.

M The agent obtains $1,000,000 in the large box.

K The agent obtains the $1,000 in the small box.

The schemas $NM_1{:}{\sim}B{\rightarrow}{\sim}P_B$, $NN_{100}{:}{\sim}P_B{\rightarrow}M_1$, and $:M_1{\rightarrow}M$ compose together to express the means–end link from ${\sim}B$ to M, contrasting with the

composition of $NM_1{:}B{\rightarrow}P_B$, $NN_{100}{:}P_B{\rightarrow}{\sim}M_1$, and $:{\sim}M_1{\rightarrow}{\sim}M$, which links B to $\sim M$ (or similarly with the just-99%-reliable N_{99} version).

And, just as in the street-crossing example, these schemas show that the schema

$*NM_1{:}B{\rightarrow}MK$

is a prejudiced-context schema, because a consequence (in the subjunctive, not necessarily causal, sense) of that schema's action B, according to the schemas above, would be P_B and in turn $\sim M_1$, which contradicts the context of $*NM_1{:}B{\rightarrow}MK$. (The schema $NM_1{:}B{\rightarrow}P_B$ could itself be construed as a prejudiced-context schema by the same reasoning. But the contradiction of that schema's context—via P_B and in turn $\sim M_1$—depends on that very schema's result P_B indeed being a subjunctive consequence of the schema's action. So we can reasonably stipulate, as a clarification of the prejudiced-context principle, that such a schema's result conditions are still considered a subjunctive consequence of the action.)

The foregoing analysis applies even if we consider a fallible simulator, as in the N_{99} case. It is then possible (though very unlikely) for the benefactor to (mistakenly) place $1,000,000 in the large box even if you will take both boxes—just as, in the street-crossing problem, it is possible (though very unlikely) to cross successfully and without collision even as dangerous traffic arrives. Even though the visible $1,000,000 thus does not imply the impossibility now of taking both boxes (similarly, even though not being about to be hit does not imply the impossibility of crossing now), it is very likely that if you were to take both boxes, the simulation would have so predicted, and the large box would have been left empty, even though in reality it wasn't (it is very likely that if I were to cross now, the traffic would hit me, even though in reality it does not), and so you should not do so (ditto). As with the opaque-box version of the problem, the expected utility of what would be the case if you were to take both boxes, or just the large one, should be computed with respect to the subjunctive probability of a correct simulation in each case (because of the means–end link from the action to the simulation result); even a modest probability of correctness (say, 0.9, or even 0.6 or less) suffices to justify the one-box choice.[8]

8. Another way to justify the one-box choice is to note that for all you know, you might be the simulated you; hence you should act in part for your (causal) influence on the simulation outcome. This view is consistent with but does not obviate the present subjunctive approach. Say the real you assumes that it is the real you. Nothing

One might challenge the alleged parallel between the street-crossing example and the transparent-boxes problem by appeal to *how you know* that the goal, or a guarantee thereof, already obtains. In the street-crossing problem, you know $\sim H_1$ in advance because you know your choice $\sim C$ in advance, whereas in the transparent-boxes problem, you know M_1 because you just see the $1,000,000. However:

- You might have an additional basis for knowing M_1—one that does depend on foreknowledge of your action. As Gibbard and Harper note (regarding the opaque box, but it applies to the transparent-boxes variation as well), you might figure out which choice you will make and what, accordingly, must have been put in the box (in the case of a reliable simulator).
- Conversely, in the street-crossing problem, you might have an additional way to know $\sim H_1$, without depending on knowing what your action will be. Suppose a rightly trusted individual assures you that you are not in fact about to be hit by traffic. Perhaps this individual is extrapolating from trillions of prior observed instances in which you, or others of safe disposition, stand waiting to cross carefully with a clear view and are never then struck by traffic, even if dangerous traffic approaches. The trusted individual need not even know whether the action of crossing occurs in any of those instances.[9]

In both problems, then, you could in principle have two simultaneous, independent bases for your foreknowledge of the actual outcome—one basis that comes from knowing what your actual choice will be, and one that is otherwise derived.

Ultimately, then, the dominance argument in Newcomb's Problem—the argument that you should take both boxes because the large box already

---

false can logically follow from that true (even if unjustified) assumption; in particular, nothing false follows as to which choice is in fact more lucrative for you to make. (The simulated you might, however, infer false conclusions from its false assumption that it is the real you.) So the one-box choice is more lucrative for you only if you cannot infer otherwise by assuming that you are indeed the real you.

9. It may seem odd that I refer to a conclusion relayed to you by a trusted third party as something *you* could thereby know confidently. But have you ever personally verified that quarks or galaxies exist, that your cells have DNA, that the Civil War took place, or that there are more than a billion people alive? A great deal of our most reliable and important knowledge comes to us by way of others' testimony.

contains either $1,000,000 or is empty, and either way, you would get $1,000 more if you were to take both boxes than if not—doesn't follow because on the contrary, if there is $1,000,000 in the large box, you would (probably) get more if you were to take just the large box than if you were to take both—because in the latter case the simulation would (probably) have so predicted and thus the large box would (probably) have been left empty (even if in fact it is not, and even though it cannot change now), regardless of whether the large box is opaque.[10] The dominance argument presumes incorrectly that the box content if you were to take both boxes (or if you were to take just the large box) *would* (in both of those cases) be the same as what that content actually *is* (and thus would be the same as what it actually will continue to be). Section 6.1's peeking-friend argument, of course, is just the dominance argument personified.

As an (albeit inconclusive) point of confirmation that the one-box choice is correct even in the transparent-boxes version, note that in that version, as in the opaque-box formulation of the problem, an individual who would choose just the large box (provided—in the transparent version—that the box contains $1,000,000) has a higher expected gain from a Newcomb's Problem encounter than does someone who would take both boxes. The former will indeed be offered $1,000,000 in the large box (or probably so, given a fallible simulation), the latter (probably) an empty large box.

Put another way, if you could contemplate the situation in advance—before the benefactor even takes the snapshot, or runs the simulation—you would hope it is the case that when the time comes to choose, you will take only the large box, even though you will already know what it contains, because otherwise the world at the time of the snapshot will be such that the simulation will probably predict your taking both boxes. If there were no rational justification for taking just the large box when the time comes, then rationality would paradoxically compel you to act in a manner contrary to how you had correctly wished (in advance) you would act under the very circumstances that you know have now arisen.

10. And similarly—but in the opaque-box formulation only—if the large box is in fact empty, you would get more money if you were to take the opaque box alone than if you were to take both boxes, because if you were to take the opaque box alone, it *would* contain $1,000,000, even though it does not and will not. (That reasoning does not carry over to the empty-large-box case in the transparent-boxes formulation, because the empty-box case is not simulated; but see sec. 6.3.)

Much the same dissonance arises in a suggestion by Rodrigo Vanegas (personal communication) in response to the transparent-boxes variation. Proposing the opposite of Nozick's peeking-friend ploy, Vanegas recommends that you *keep your eyes closed*, transforming the problem back to an opaque-box problem, allowing you to take just the large box and reap $1,000,000. Otherwise, the evidentialist argument goes, what you see in the large box (regardless of whether you see $1,000,000 or an empty box) would rationally compel you to take both boxes, and the large box would thus be empty (given a correct prediction of your rationally compelled choice).

From an evidentialist perspective, ignorance of the large-box content may indeed be necessary in order to choose correctly for the sake of that content. But the present theory of subjunctive means–end links lets you make the more lucrative choice even if you do look at the box (or even if the rules of the encounter were to require you to look). As in the street-crossing example, you can rationally act for the sake of what would be the case if you were to make one choice or another, even if you already know what actually *will* be the case (regarding your choice and its outcome).

The fatalist intuition that it is futile to act for the sake of an already-known, already-inalterable goal can be induced by focusing on the known prior guarantee in the deterministic street-crossing scenario. But the intuition asserts itself far more insistently in Newcomb's Problem, especially the version with transparent boxes. I suspect the disparity comes from the differing obviousness of the present-time, already fixed, already known guaranteeing conditions in the various problems:

- With the transparent large box, the guaranteeing condition is simple, concrete, and plainly visible.
- With the street-crossing example, the condition is complex and abstract, and we are easily oblivious to it. (In fact, people typically deny it outright by rejecting determinism and positing free will.)
- The original version of Newcomb's Problem is intermediate: there, the guaranteeing condition is simple and concrete, but not visible.

In real-world situations, we seldom if ever observe an obvious physical condition that we know depends subjunctively on a forthcoming choice.

Instead, in the obvious cases, an already observed, already inalterable physical condition is indeed beyond the reach of our choices.[11]

Conceivably, then, even if our machinery has a built-in prejudiced-context principle, we might develop a belief that it is futile to act for the sake of an already inalterable outcome. That belief might arise in part as an overgeneralization from the obvious cases—much as, say, the overdue-heads belief mentioned in section 5.2 might arise as an overgeneralization from other kinds of situations where an overdue expected arrival is indeed especially likely to occur soon. Just as our intuitions about inductive reasoning can be improved by augmenting our built-in induction machinery with a correct explicit theory, so too can our intuitions about choices and goals benefit from an explicit theory of means–end relations—in particular, perhaps, a theory that includes something like the prejudiced-context principle.

In Newcomb's Problem with a transparent large box, *seeing* a virtually certain prediction of your choice—and seeing the goal state itself (or a guarantee thereof)—makes it impossible to ignore or deemphasize that the choice and the goal state are already established. The seeming conflict between choice and determinism intrudes before our very eyes, exposing any lingering allegiance to the fatalist intuition.

But the visible goal state in the transparent-boxes problem is no more preestablished than goal states always are, given determinism. It is just more *blatantly* preestablished.

6.2.3 Or What If the Answer Is Not Built In?

Although I have argued above that both the explaining-away and prejudiced-context principles should be built into well-designed choice machinery (as well as being principles that an intelligent agent might

11. There is one important exception. The next chapter argues that others' observed (or reliably predicted) cooperative choices in Prisoner's Dilemma situations may depend subjunctively (even if acausally) on one's own forthcoming choices. But (unless we agree with that argument) we do not ordinarily recognize a clear, uncontested subjunctive link in such situations. Instead, evolution may have rigged us with inclinations (to empathy, to tit-for-tat behavior, etc.) that roughly mimic a recognition of the subjunctive means–end link in those situations; and we may invent imaginary causal links (karma, afterlife incentives, etc.) to stand in for the acausal subjunctive link. The next chapter elaborates these points; see also section 6.2.3 just below.

explicitly analyze, as you and I are now doing), there is good reason to suspect that the prejudiced-context principle may not in fact be built into our own machinery.

Evidence for the explaining-away principle being built in (at least if the overall schema framework is a roughly accurate portrayal of part of the machinery) is that in paradigmatic situations such as the street-crossing scenario, using schemas without recourse to explaining away would leave an unresolved conflict about whether, say, dangerous traffic would vanish if I were to cross in front of it (secs. 5.5 and 5.6), in the choice-supporting sense of *would*. In fact, though, we effortlessly find it intuitively obvious that crossing would have no such consequence (even if we feel at least partly persuaded of acausal consequences in certain other situations, such as the hand-raising scenario or even Newcomb's Problem); a built-in explaining-away principle would account for that obviousness. (The non-vanishing-traffic outcome is obvious even to people who have never explicitly contemplated anything resembling the explaining-away principle. Hence, if invoking the principle is indeed responsible for the obviousness, then the principle must be built in.)

The prejudiced-context principle, in contrast, seems to lack such salutary influence on our intuitions. I believe the account I have presented here; yet if I imagine myself in the transparent-boxes situation, I still feel the strong tug of the intuition to take both boxes. To bolster the one-box prescription, I need to argue with myself, reiterating the explicit analysis, focusing on the two constituent means–end links (from my choice to the snapshot-time universe state, and from there to the box content) and on the better average outcome for one-box choosers in transparent-boxes encounters.

More strikingly (since, after all, my controversial one-box prescription could just be wrong), the prejudiced-context principle seems unable to fully suppress the inalterability-implies-futility intuition even in mundane situations like the deterministic street-crossing scenario with foreknowledge of the noncollision outcome (sec. 6.2.1). On the contrary, people often insist that determinism would indeed make choice futile even in such clear-cut situations. Accordingly, many reject determinism and invent an incoherent "free will" to preserve a sense of efficacy of their actions. Even those who explicitly disavow free will may still need to pretend otherwise in order to salvage the feeling that choices matter (for example, Minsky advocates such a subterfuge in *The Society of Mind*, p. 307). When I

reflect that the future and past alike sit immutably in spacetime, I do feel an uncomfortable challenge to the notion that my choices make a difference, even in the most clear-cut instances.

This persistent conflict of intuitions, in contrast with the seemingly self-evident resolution of the conflicting evidence about the prospect of vanishing traffic, is just what we would expect (as discussed in sec. 5.6.4) if indeed the principle in question were *not* built into our choice machinery. And insofar as the conflict resolved by the prejudiced-context principle does not arise in ordinary situations *except* given the (modern and sophisticated) acknowledgement of determinism (but see n. 11 just above), there may have been no selective pressure for evolution to have built the principle into our cognitive architecture.

In the absence of a built-in prejudiced-context principle, the best we can do is to explicitly convince ourselves that some such principle *should* be invoked, since the alternative would have us conclude absurdly that all actions are futile, even in routine situations (were we not to dodge that absurd conclusion by pretending that genuine choices cannot be predetermined and even foreknown). Unfortunately, we are then left with a free-for-all between, on one side, our explicit theory, together with those of our means–end intuitions that comport with that theory (intuitions that prescribe appropriate actions in routine situations); and on the other side, our still-active contrary intuition that inalterability implies futility (especially in the case of foreknown outcomes). This conflict is then subject to whatever heuristics our machinery may use to muddle through such confusions (recall n. 22, chap. 5, and the subsequent discussion of explicitly held beliefs versus built-in machinery), with no built-in basis to automatically, decisively resolve the contest.

Thus, the absence of a built-in prejudiced-context principle (from a mechanism that is otherwise along the lines sketched here) could leave us with intuitions that recognize means–end links (including some acausal ones), but without robust immunity to the contrary intuition that given determinism, all putative means–end links are bogus (even causal ones). That may indeed be where we stand. Perhaps that helps explain why the reconciliation of choice and determinism has been such a stubborn philosophical problem.

This hypothesis amounts to a much less pessimistic version of Chomsky's speculation (Antony and Hornstein 2003) that certain important

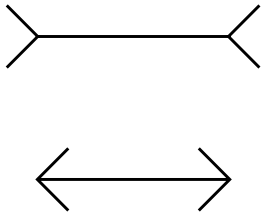


Figure 6.1
The Müller–Lyer illusion.

puzzles—such as how we could exhibit apparently free choice, despite the physical constraints on our constituent parts—may turn out to be not just hard problems, but rather eternal *mysteries*: problems whose solutions our brain architecture just does not equip us to grasp, much as a rat's brain can solve mazes but not factor numbers. But the present hypothesis is that we *can* come to comprehend a correct explicit theory of choice and means–end relations—perhaps a theory including something like the prejudiced-context principle—and thereby be influenced to choose as the theory prescribes, even if an architectural limitation (e.g., the omission of a prejudiced-context principle from our built-in choice machinery) keeps the contrary inalterability-implies-futility intuition in force as well.

Analogously, an optical illusion, too, can stay in force even when we explicitly know better. For instance, measurement reveals that the two horizontal lines in figure 6.1 are of equal length, though the impression persists that the upper one is longer. Despite that illusion's continuing distraction, we can base our considered decisions on our knowledge of the lines' equality. We face a curiosity, not an insurmountable mystery.

6.2.4 Contrast: Kavka's Toxin Problem

Simon Blackburn (1998) discusses a problem by Greg Kavka (1983), the *toxin puzzle*—which, as Blackburn notes, is structurally close to Newcomb's Problem with transparent boxes, with a large, visible reward bestowed if and only if you are predicted to later forgo a small reward—and advocates the equivalent of the one-box choice. (The small reward in Kavka's scenario consists of not drinking a mildly noxious potion; hence the puzzle's name.)

But in Kavka's variant, the predictor tells you what the game is prior to making the prediction, and then reads your mind accurately enough to predict what your eventual choice will be, offering or withholding the large

reward accordingly. Blackburn argues that you should "cultivate the disposition" to forgo the small reward (for the benefit of that disposition's conventionally causal influence on the still-future mind-reading), and to refrain from changing your mind even after the large reward is obtained (or else, by assumption, your resolve would not have been firm enough to convince the predictor). Still, if your choice process operates as always when it comes time to take or forfeit the small reward, Blackburn does not, I think, explain what reason you then have to choose to abide by your earlier resolution.

But perhaps the choice process does *not* then work as always (that is, by comparing what would be the case if one or another action were to occur, and taking the action whose consequence is preferred). Perhaps instead your cognitive machinery includes a resolution-enforcement or willpower module that engages if you resolve to later forgo the small reward, and that subsequently disables your preferred-consequence-selection process, or at least competes against that process for control of your actions. (It is almost obligatory here to invoke the metaphor of Ulysses binding himself to a mast to restrain himself from later choosing to follow the Sirens' call.) To the extent that the subsequent choice process might thus be overridden, your making a resolution might literally prevent you from having a subsequent choice in the matter—not merely because the action is inexorably predetermined, but rather because the choice machinery is prevented from functioning when it comes time to act, or at least prevented from controlling which action occurs.

However, circumventing the very choice process would merely circumvent, rather than solve, the choice paradox—the paradox as to what you later should choose if you do then still have a choice. (Suppose we amend the rules to forbid anything that overrides the subsequent preferred-consequence-selection process—such as engaging an internal resolution-enforcement module, if one exists—on pain of forfeiting the entire reward.) And empirically, our resolve (whether or not it is in fact implemented with help from a separate enforcement module) has only limited efficacy against subsequent flip-flopping. To the extent that the subsequent choice process is *not* overridden—to the extent that it still exerts control over your actions—it still needs a basis on which to calculate that what would be the case if you were to forgo the small reward is preferable to what would otherwise be the case (if, in Kavka's problem, the resolution-abiding choice is to be vindicated). Blackburn does not offer such a basis.

Moreover, in Newcomb's Problem, unlike in Kavka's variant, the entire problem is presented to you not in advance of the mind-reading (the snapshot and simulation), but only afterward when your large reward has already been secured. (The Newcomb's Problem snapshot and simulation might even have been conducted before you were born.) By the time you first confront the problem, your then-cultivated disposition can exert no causal influence upon your large reward, for the large reward has already been established before you make any resolution. Hence, Blackburn's proposed solution would not successfully defend the one-box choice in the transparent-boxes problem.

6.3 Newcomb's Problem with a Dual Simulation

As a final exercise, to make the scenario even more extreme (as though it were not radical enough already), we can contemplate yet another transparent-boxes variation of Newcomb's Problem. In this new scenario, the predictor conducts *two* simulations, one (as in the previous scenario) showing you presented with $1,000,000 in the large box, the other simulation showing you presented with an empty large box. (This added twist is not gratuitous; in particular, it contributes to the discussion below in sec. 7.2.2 of ethical foundations and the Rawlsian veil of ignorance, and to the discussion of other aspects of ethical theory in secs. 7.2.2 and 7.3.1. However, a reader who has reached the saturation point can defer this section without much loss of continuity.)

In this new variation, the benefactor places $1,000,000 in the large box if and only if *both* simulations show you taking just the large box. Both simulations are highly (but not perfectly) reliable in their predictions of what you would do if you in fact encountered the specified large-box content. As always, the benefactor informs you accurately of the rules.

Let us redefine the predicate P_B to be true if and only if the snapshot-time state of the universe is such that (according to correct physical laws) you will take both boxes if you are presented with an empty large box, or if you are presented with $1,000,000 in the large box, or in both cases; that is, P_B is false if and only if you will take just the large box unconditionally, regardless of whether it is empty. Assume that if $\sim P_B$, there is a 0.99 chance that both simulations accurately predict your one-box choice, resulting in $1,000,000 in the large box. Assume further that if P_B, there is likewise a

0.99 chance that the simulations accurately show you taking both boxes in at least one of the two cases, resulting in an empty large box.

I claim that in this variation of Newcomb's Problem, you should still take just the large box—even if it is empty. The argument is as follows.

If and only if you were to take just the large box (unconditionally, regardless of its content), then $\sim P_B$ would be the case. If and only if $\sim P_B$, the dual simulation would (very probably) predict your taking just the large box (in both branches of the simulation), and thus there would be $1,000,000 in the large box (even if, in fact, you can see that it is empty, and you know it cannot now change). Thus, using the same subjunctive reasoning as in the previous version of the problem, we can conclude (correctly, I claim) that you should take just the large box even if it is empty—provided that the dual simulation is reliable enough that the expected utility of $\sim P_B$ is greater than that of P_B—that is, the expected utility is greater if (according to physics) you are going to take just the large box, regardless of its content, than if you are going to take both boxes, either unconditionally or as a function of the large-box content. (The presumed 0.99 accuracy easily makes that expected-utility calculation favor $\sim P_B$, which in turn favors taking just the large box.)

By ensuring that $\sim P_B$ is the case—by ensuring that the snapshot-time state is such that you will take just the large box (unconditionally, whether it is empty or not)—you incur the reward of $1,000,000 in the (much more probable) case in which the simulation does not err. Accordingly, the one-box prescription is supported by the same points of confirmation as in previous versions of Newcomb's Problem:

- Taking just the large box—even if it is empty—is what you would correctly wish in advance (prior to the simulation) that you will do, because if that is what you will do, the simulation will probably so predict, and you will then obtain $1,000,000.
- Someone who would take just the large box (even if it is empty) reaps a larger payoff, on average, from dual-simulation Newcomb's Problem encounters than someone who would not.

By using an expected-utility calculation with regard to the simulator's reliability, you effectively act for the sake of the entire probability-distribution of outcomes as to whether the dual simulation simulates your choice(s) accurately or not. But what if you can deduce, before making your

choice, whether the simulation was in fact correct on this occasion? If you subscribe confidently to the present analysis, you may thereby know in advance that you will choose just the large box. Hence, given the visibly empty large box, you can conclude that the simulation must have erred this time. In that case, would you still be right to choose the large box alone—a choice that is motivated by the probability that the simulation predicts accurately? If not, then the above analysis is not correct.

The possibility of deducing that the simulator erred thus poses an additional challenge, threatening to undermine the subjunctive reasoning above: *given* that a simulation error has occurred, it is no longer the case that if P_B were false, there would be $1,000,000 in the large box—just the opposite, in fact. The rest of this section addresses this additional challenge. The role of this analysis is a bit subtle, and should be made explicit:

- Again, you can deduce simulation-error when you see the empty large box, *if* you know that you will unconditionally choose just the large box (as the present account advocates). But if instead you know you will choose both boxes, contrary to what the present account advocates—or if you just don't know either way—then you *can't* deduce that the simulation erred. In that case, the additional problem under discussion now does not even arise—the problem of justifying a utility calculation that is based on the (large) probability of a correct simulation, even when you can deduce that the (unlikely) event of a simulation failure has in fact occurred.
- If the present account is right, though, then that account has to be defensible even if you are justly confident that you will always take just the large box, as the present account advocates; and therefore, the account has to be defensible even if you can deduce that the simulation erred in this instance. That is, even given that you will take just the large box, you must still (if the present account is right) be able to defend the conclusion that you *should* take just the large box. It is that defense that the remainder of this section undertakes.

I maintain that you should indeed still choose just the large box. A preliminary observation is that the two points of confirmation above still hold in this case: taking just the large box (even if it is empty, and even if you can thereby deduce that the simulation erred) is what you would correctly wish in advance (prior to the simulation) that you will do; and someone who would do so reaps a larger payoff, on average, from these encounters than someone who would do otherwise.

But even if we accept the acausal means–end link in the case of a transparent box containing $1,000,000, it seems highly counterintuitive, in the case of an empty large box, that you should act for the sake of a probability-distribution of outcomes (as to the simulation's accuracy) even when the *actual* outcome (the simulation's error in this instance) is known to you. As just noted, the subjunctive argument seems no longer to go through in the case of deducible simulation error.

Perhaps the most vivid illustration of the argument for still taking just the large box (even though it is empty and you can deduce that the simulation erred) arises if the simulation's accuracy depends on a quantum coin toss—that is, an event with a quantum superposition of outcomes, as discussed in chapter 4. If we accept Everett's relative-state view (as argued for in chap. 4), we obtain a superposition in which both outcomes actually occur, in different branches of configuration space (but, by stipulation here, the accurate-simulation outcome is much more heavily weighted than the erroneous-simulation outcome). Thus, the acausal means–end link to the snapshot-time past state chooses between two configuration-space situations:

- In one situation, you take just the (empty) large box. Hence, the snapshot-time past state is such that you will subsequently take just the large box if it is empty. Then—provided that you will also do so if the large box is not empty—P_B is false, and the dual simulation results in your getting $1,000,000 in the region of configuration space that receives most of the quantum squared-amplitude, namely the region where a correct prediction occurs; but the large box is empty in the configuration-space region that receives a lesser portion of the quantum squared-amplitude, the region where an incorrect prediction occurs. (It is in this lesser-weighted part of configuration space that you take just the empty large box.)
- In the other situation, you take both boxes. Hence, the snapshot-time past state is such that you will subsequently take both boxes (at least if presented with an empty large box). Then, P_B is true; the dual simulation results in your getting an empty large box in the greater-weighted configuration-space region, where a correct prediction occurs; but the box contains $1,000,000 in a lesser-weighted region, where the simulation errs. (It is in the greater-weighted part of configuration space that you take both boxes.)

If the probability of a correct simulation is at least modestly high, then you fare better (averaged across all the quantum squared-amplitude in configuration space) in the first situation (where P_B is false) than in the second (where P_B is true). Thus you do better overall if you take just the large box, even when it is empty.

But what if the simulation's accuracy hinges on an ordinary random event, not a quantum coin toss (or what if Everett's interpretation of quantum events turns out to be wrong)? Then it is not literally the case that both probabilistic outcomes (a correct simulation and an incorrect one) actually occur (though the configuration-space distribution of outcomes remains a good illustration of the probability-distribution of outcomes). Nonetheless, I claim that you should still make your choice as though the entire range of outcomes were actual (weighted by their respective probabilities)—for the same reason that you should do so (thereby using a conventional expected-utility calculation) in the ordinary case where the outcome of a random event is not already deducible.

Let us digress briefly to address what that reason is. Suppose, for example, that I offer you $10 if the next toss of a fair coin comes up heads, provided that you pay me $1 to play the game. Why is that a good bet for you to accept? How does the desirability of receiving $10 imply that a 50 percent chance of receiving $10 is worth more than a definite $1—or even that it is worth anything at all?

This is another foundational question whose answer cannot be deduced from scratch, on pain of circularity. Recall the discussion in section 5.2 about the similar status of inductive reasoning and of means–end links. You can observe that inductive reasoning has worked until now, but it is circular to conclude (inductively) that it will therefore continue to work. Likewise, you can observe that if you were to construe subjunctive links as means–end links, you would better achieve your goals than if you were to use various alternative construals (evidentialist, exclusively causal, fatalist, etc.). But it is circular to adopt the subjunctive construal on the basis of that (subjunctive) benefit.

A similar point applies now to the use of expected-utility calculations to ascribe (attenuated) utility to probabilistic outcomes. (The similarity is unsurprising, since decisions grounded in expected utility combine probability with means–end links.) You can observe that if you were to treat

probabilistic outcomes as having the utility of already certain outcomes, but attenuated by their respective probabilities, you would almost certainly fare better in the long run than if, say, you attributed no utility to probabilistic outcomes. But to attribute such utility for that reason would be circular, because we need the very principle under consideration to justify attributing any utility to that not-quite-certain long-run outcome.[12]

Fortunately, an analytical choice machine that is built to perform inductive generalization, or to use means–end relations, simply does so without needing an explicit reason (in the sense of rationale, as opposed to cause), just as the heart beats without needing an explicit reason (rationale) to do so. And this is so too for attributing (attenuated) utility to each of a range of probabilistic outcomes, at least in straightforward situations like the coin-toss bet where schemas connect a concrete action (accepting the bet) to a range of possible outcomes (the then-observed coin toss can come up heads, or tails); the schemas' reliabilities then can be approximated by empirical tabulation over a number of trials, and used in built-in utility calculations, as outlined in section 2.4.1.

In less straightforward situations, the built-in machinery may need to acquire an explicit knowledge of probability, including the sort of explicit expected-utility calculations that we sometimes consciously perform. But in contrast with the hardwired utility calculations in straightforward cases, any explicit calculations we perform again raise the foundational question: what then makes us *care* about maximizing an explicitly calculated expected utility?

The delegated-value machinery proposed in section 2.4.1 may offer an answer. Recall the gist of that idea: the choice machine should delegate a kind of strategic value to a state (such as having money in the bank) that may lack any current tactical, instrumental value (suppose there's nothing you need to buy at the moment), but which is often of instrumental value,

12. More accurately, we need at least a special case of the principle under consideration. If we at least could start with the principle of attributing near-full utility to near-certain outcomes, we could use the law of large numbers to bootstrap from there to attributing attenuated utility to arbitrarily uncertain outcomes (because with sufficiently many repetitions of an event that has a given probability distribution of outcomes, the law of large numbers tells us, with near certainty, that the set of actual outcomes closely approximates the probability distribution that each constituent event had).

and which—if you wait until you currently need it to try to obtain it—will not be readily obtainable (if your bank account is depleted, it may be difficult to restore it promptly when the need does arise). Roughly speaking, if you wait until you need the state in question, you've waited too long, so you want to pursue that state regardless of its current need, as though it were valuable in and of itself; hence, delegated value.

My speculation, then, is that explicit representations of expected utility would qualify to receive delegated value (positive and negative) from the possible outcomes over which the utility is calculated. Strategically pursuing the explicitly represented expected utility of an action (e.g., representing the expected value of having placed a given bet) puts you in a position to reap the benefit of the positively valued outcomes; but you cannot successfully pursue that benefit tactically, by betting only if a positively valued outcome occurs, because by the time you know what outcome occurs, it is too late to place the bet.

Value delegated to expected-utility representations might also help answer a reasonable question about the quantum-randomness illustration above: why should you care what happens to versions of yourself in remote configuration-space branches? (This question can be formulated in terms of identity—are other versions of you "really you"?—but asking why you should care—or asking what machinery might make you care—is more substantive than trying to gerrymander the border between "you" and "not you.") Chapter 4 argued that from a subjective standpoint, by virtue of distributions manifested in cumulative records or memories, configuration-space multiplicity is tantamount to a probability distribution. Hence, the same delegated-value considerations apply when dealing with a configuration-space distribution as when dealing with a probability distribution.

In contrast with the coin-toss bet above, suppose I offer you a (rather odd) opportunity to bet *retroactively* that heads came up (using the same fair coin as before), when you can already see (or else deduce from some indirect evidence) that the toss came up tails this time. If the whole probability distribution of outcomes should be valued (as I claim), then why should you not still be willing to bet on heads? Of course, to do so would be absurd. But conversely, if the whole distribution should not still be valued once the actual outcome is known, then how can it make sense (as I claim)

still to value the whole distribution in the empty-box scenario, when the losing outcome (i.e., the unlikely simulation-error) is likewise already deducible?

In fact, an unconditional choice to bet retroactively on the coin toss—regardless of the observed outcome—is indeed better than an unconditional choice not to bet (assuming that the retroactive bet is offered to you regardless of whether the outcome is heads or tails), as can be seen by considering the probability-distribution over the possible coin-toss outcomes given either the decision to bet or to not bet. However, the retroactive-bet offer makes a third alternative available to you: rather than betting unconditionally or not-betting unconditionally, you can bet on heads if and only if the coin toss did come up heads. That choice about the retroactive bet (obviously unavailable if instead you must bet in advance) is clearly the best of the three. And making that third choice is still consistent with valuing the whole probability-distribution of coin-toss outcomes: the third approach gives you the best average outcome across the distribution.

Returning now to the empty-box scenario, the salient point is that even though the bet here (as to the simulation's correctness this time) is placed retroactively (by means of your one-box or two-boxes choice), and you can already deduce that the simulation erred this time, there is a crucial difference here from an ordinary retroactive bet. For here, the choice of whether to take both boxes has a (subjunctive, not causal) side consequence about whether P_B is the case. And with regard to P_B, the equivalent of the retroactive coin-toss bet's third approach—namely, the choice to have P_B be the case if and only if the simulation did err—is unavailable here. That approach is unavailable (even if you can deduce that the simulation erred before you choose one or both boxes) because there is no way, under the stipulated rules, for the truth of P_B to depend on the forthcoming simulation-accuracy (which, by stipulation, is 0.99 whether P_B is true or not).

So if you were to choose both boxes (and thus ensure that P_B is true) when the simulation erred and the large box is empty, you would also thereby ensure that P_B would be true even if (more probably) the simulation had *not* erred in this instance. This subjunctive consequence contrasts with an ordinary retroactive bet: rejecting the coin-toss bet if you see or deduce that the toss came up tails would not thereby ensure that you would have rejected the bet if instead the toss had come up heads. If you were to

choose both boxes when the large box is empty, you would fare worse, averaged over the probabilistic distribution of simulation-accuracy outcomes, than if P_B were false and you were thus to take just the large box.

Thus, given the dual simulation, you should take just the large box, whether or not it is empty. And again, as a point of confirmation, someone who would do so fares better, on average, in dual-simulation transparent-box Newcomb's Problem encounters than does someone who would, say, take both boxes in the empty-box case.

6.4 Summary

To understand how genuine choice could be mechanical—to reconcile choice with determinism, or even with approximate determinism—we must confront the compelling fatalist intuition that it is futile to act for the sake of that which our action cannot alter—the intuition that inalterability implies futility. Contrary to that intuition, we rationally act for the sake of what would be the case were we to do one thing or another, in the choice-supporting sense of *would*. In such cases, we can say there is a means–end relation (or informally, a means–end link) between our action and our intended goal. An action so taken is a choice, and its success does not involve changing anything from what it is already predetermined to be.

Even foreknowledge of an outcome does imply that we have no choice about the outcome. Foreknowledge only seems to preclude choice insofar as foreknowledge is a testament to inalterability, and insofar as we misconstrue inalterability as proof of futility. Genuine choice requires neither the alterability of an outcome, nor prior ignorance or uncertainty about the outcome.

In the previous chapter, contemplating the operation of a simple choice machine demonstrates that choice is a particular mechanical process that, like any other such process, is no less real for having been predetermined. Contemplating the choice of a distant-past state (e.g., in the hand-raising example) shows that, once inalterability is no longer deemed a prohibitive obstacle, we can be seen to have a choice about some aspects of the world over which we lack any causal influence; that is, there can be an acausal means–end link from our actions to some aspects of the world's state.

Newcomb's Problem simply harnesses our ability to choose some such aspects of the world, using an acausal link to the snapshot-time past state; that state then serves (via a causal path) as a subgoal to the goal of the eventual reward. But of course, as mentioned in section 5.4, an acausal means–end link is not some strange kind of "force." Rather, a means–end link, causal or otherwise, is just an abstract relation between an action and a goal—a relation such that the desirability of the goal rationally motivates taking the action (other things being equal).

Newcomb's Problem distills the challenge posed by deterministic choice. If inalterability does not imply futility, then being unable to alter the box's content does not necessarily undermine the desirability of acting for the sake of its content being one way or another. The transparent-boxes version goes one crucial step further. If an outcome is already determined anyway, then already knowing what the outcome will be—or even literally *seeing* what it will be—does not further undermine an action's efficacy with respect to that outcome. If it can make sense to act for the sake of the unknown, inalterable box content, the same holds even if instead that inalterable content is already known.

Making both boxes transparent makes the seeming conflict between choice and determinism especially vivid, without changing its essential character. The concluding remark of section 6.2.2 bears repeating:

> The visible goal state in the transparent-boxes problem is no more preestablished than goal states always are, given determinism. It is just more *blatantly* preestablished.

If the choice–determinism conflict is not resolved in its most blatant form, then it is not resolved, but rather just partly concealed. In a universe such as ours, with no flow of time (despite our impression to the contrary, as discussed in chap. 3), all spacetime is a static, already sealed box. That the box content is already inalterable (and often already foreknown) must not necessarily prohibit acting for the sake of that content being one way or another, if choice is indeed rational in such a universe.

Whereas much of the literature construes arguments against evidential means–end relations as supporting exclusively causal means–end relations and vice versa, I propose here an intermediate approach. I construct a subjunctive sense of means–end relations—a choice-supporting sense of what

would be the case if this or that action were taken—broad enough to include some acausal evidential relations, but narrow enough to exclude others.

The machinery I sketch, in this chapter and the previous one, for recognizing means–end relations uses schemas that express correlations among specified conditions. A preliminary presumption of conditional independence from all other conditions can be (unremarkably) superseded by an exception-override provision. An explaining-away principle can selectively defeat the preliminary presumption that a schema's evidential relation is also a means–end relation, that its conditional probability is also a subjunctive probability. And finally, the prejudiced-context principle can defeat the same presumption with regard to a schema whose context depends subjunctively on the action itself.

The means–end-recognizing machinery here is proposed tentatively. I offer no proof that the explaining-away and prejudiced-context principles successfully do the work required of them. Instead, I present only some plausibility arguments focused on paradigmatic situations.

Accordingly, I do not expect to find that the details of the proposed machinery are complete and correct. But I am hopeful about the merit of the general approach that those details illustrate—the methodology of using thought experiments that presume determinism and zero-probability idealizations, and the attempt to derive candidate means–end links from correlations among events in actual situations, then winnow the candidate links in part by the deference of some links to more-general explanatory links, and in part by deference of a link whose context depends (in the appropriate subjunctive sense) on the very action under consideration.

The foregoing approach tries to justify the proposed means–end recognizing machinery by appeal to mundane situations (e.g., the street-crossing scenario) that help to isolate and examine the relevant principles. These are the sort of situations that our cognitive machinery must have evolved to deal with. Thus grounded, the principles embodied in our means–end machinery can then be applied to esoteric or controversial scenarios, such as Newcomb's Problem or (in the next chapter) the Prisoner's Dilemma.

I speculate that the one-box intuition in Newcomb's Problem (and the corresponding procooperation intuition in the Prisoner's Dilemma) reflects our choice machinery's recognition of an acausal means–end link—a correct recognition, by the present account, facilitated by schemas' use of

correlations and of the explaining-away principle. (More tentatively, I speculate that the *absence* of a prejudiced-context principle from our built-in choice machinery might help account for the contrary intuitions in those problems—and also for the perceived futility of *all* choices, given determinism, even in mundane situations.) As too is the case with inductive reasoning, our built-in means–end recognition can be recapitulated and extended by our explicit analysis of the machinery's underlying means–end-recognizing principles, as attempted in this chapter and the previous one.

7 Deriving *Ought* from *Is*

Why behave ethically or fairly or with respect for others? Why not lie, cheat, steal, and kill to the extent that you can profit from it and get away with it? Especially if people are just configurations of inanimate atoms, why have any more ethical regard for a person than for any other collection of atoms? Is there a genuine mistake of reasoning committed by those whose conduct is entirely amoral,[1] who have no aversion to harming others except to the extent that their own interests would thereby suffer too?

Most of us have the overwhelming intuition that there is indeed something irrational about such conduct, some important truth that such a person does not get. But through the centuries, it has proven difficult to explain just what the amoral person's error of reasoning consists of. It is easy enough to proclaim some unspecified error, or to proclaim that there are specific moral principles (that it is wrong to kill, for instance) without being able to demonstrate that those principles are correct. But the mistakenness of total amorality (or of less extreme variants) is not legitimately demonstrated by unfounded proclamations. I strive in this chapter to defend the intuition that amoral conduct is indeed mistaken.

The previous chapter's discussion of mechanical choice laid the groundwork for exploring next how such choice can be subject to rationally derived ethical constraints. The argument so far is that choice is compatible with determinism (which is important regardless of whether this universe does happen to be fully deterministic, because the universe is at least deterministic *enough*, in many mundane circumstances, for the principles of deterministic choice to carry over). Choice is a matter of acting for the

1. As noted at the outset, I use the terms *moral* and *ethical* synonymously to designate matters of right and wrong.

sake of what would then be the case, which can differ from what in fact is (or was or will be) the case. Pinning down the choice-supporting sense of *would* by looking at mundane situations shows us how to resolve some paradoxes of subjunctive reasoning that arise in esoteric thought experiments such as Newcomb's Problem.

We are now in a position to address some less esoteric thought experiments—namely, Prisoner's Dilemma situations, which exemplify some of the puzzles of cooperative, altruistic behavior, without presupposing any fantastic simulators or the like. Prisoner's Dilemma situations are in one sense quite ordinary—situations of similar structure arise all the time in real life. They seem peculiar only in light of the proposal that in the Prisoner's Dilemma, there is in fact a means–end link from the action of cooperating with another, to the goal of another's cooperation toward oneself, even in situations where cooperation cannot *cause* such reciprocity.

The existence of such a means–end link—analogous to the acausal means–end link in Newcomb's Problem—is just what this chapter proposes. This account suggests a foundation for ethics—a way to show why it is rational to treat others well, even when doing so causes no net personal benefit (or even causes net personal harm). For such a foundation to be solid, it cannot simply take for granted any fundamental ethical premise, and the argument here is that no such premise is needed. Instead, I claim, it turns out that treating others well is rationally motivated by a subjunctive means–end link to others' reciprocity, even when there is no causal link. Others would have reason to treat you well only to the extent that you yourself were to have reason to treat others well, in the choice-supporting sense of *would*.

7.1 From Newcomb's Problem to the Prisoner's Dilemma

Creating a Newcomb's Problem situation would not necessarily require a fantastic mechanism for taking an accurate particle-by-particle snapshot and running a faster-than-reality simulation using that starting point. As Dennett (1987) has pointed out, informal or *folk-psychological* knowledge of others' behavior is often as reliably predictive as the best scientific knowledge.

For example, as Hofstadter (1985) notes, if I ask you to solve an easy arithmetic problem—say, adding a pair of three-digit numbers—I can pre-

dict your answer with reasonable reliability. I need not anticipate what each of your atoms or even each of your neurons does when you add the numbers. I need not even use the same addition algorithm as you; we might process the digits in different orders, for instance. Still, by adding the numbers myself—by solving the problem that I know we are both competent to solve—I can predict your own answer. Unlike the science-fiction simulation, though, this high-level simulation depends on our mutual competence. If it is not reasonably assured that you and I are both able to solve the problem correctly, then I may have no way to guess what answer you will arrive at.

Suppose the previous chapter's analysis of Newcomb's Problem turns out to be correct, and imagine that the analysis someday becomes so widely accepted and uncontroversial that among educated persons, competence to solve Newcomb's Problem as above can be confidently presumed, just like arithmetic competence. At that time (but not necessarily before), I can adequately simulate your thought process in Newcomb's Problem simply by solving the problem myself; I can thus set up a Newcomb's encounter using that folk-psychological simulation of your choice deliberation, in place of an atom-by-atom simulation. I would place $1,000,000 in the large box (whether it is opaque or transparent), confident that you would make the correct choice and take the large box alone.

The simulator-based Newcomb's Problem relies on an acausal means–end link that says:

- Given the Newcomb's Problem setup: if and only if the chooser were to take both boxes, the snapshot-time past state of the universe would be such that (according to physics) the chooser will take both boxes.

This is a past-predicate link, as discussed in the previous chapter. It composes in turn with a causal link:

- If and only if the past state were thus, the simulation would predict the action of taking both boxes, and the predictor would place no money in the large box.

In contrast, the human-competence version of the problem relies on a different sort of acausal link—not a past-predicate link, but rather one that says:

- Given that the chooser—faced now with a problem that lies within its competence—carefully solves the problem and acts accordingly: if and

only if the chooser were to take a given action, then that action would be the one regarded as correct by a competent choice process when faced by that problem (presuming the problem is such that there is a unique correct action).

That link composes in turn with a link that says:

- If and only if the action regarded as correct by a competent choice process were thus, then another competent chooser contemplating the same problem would regard that action as correct, and would accordingly predict that a competent chooser will take that action.

There is of course nothing you can do that can change what choice a competent choice process arrives at in the actual situation; rather, what that choice is is logically inherent in the situation itself. But similarly, there's nothing you can do to change the box content in Newcomb's Problem, or (in the previous chapter's street-crossing scenario) to change the present universe-state's assurance of your safety. Still, a subjunctive link does not require being able to change anything—as argued in the previous chapter, inalterability does not imply futility. Rather, a subjunctive link requires only the right kind of contrast between situations in which a given action occurs, and situations in which it does not. Such a contrast implies a correlation that—in conjunction with some additional criteria, as given, for example, by the explaining-away principle—implies a subjunctive link.

With the folk-psychological simulation, the requisite contrast involves a range of different choice problems that fall within the agents' mutual competence. For any such problem, if one competent agent concludes that a given action is correct and acts accordingly, another competent agent contemplating the same problem arrives at the same answer. Thus, the subjunctive link here is not via what the past universe-state would be (hence what snapshot-input the simulator would receive), but rather via what choice a competent agent would regard as correct (hence what the predictor would so regard). One consequence of this difference from the atom-by-atom simulation is that the folk-psychological Newcomb's simulation (like the folk-psychological arithmetic simulation) might not work except to the extent that both parties are competent to solve the problem correctly.[2]

2. Another consequence, addressed just below in the discussion of the Prisoner's Dilemma, is the need for an alternative argument for why the explaining-away principle of section 5.6.3 does not override the acausal link.

Consequently, if the account here of Newcomb's problem is indeed correct, the account's present obscurity implies that there are likely very few predictor-chooser pairs today for whom the requisite means–end link exists, mediated by a folk-psychological simulation. It is also difficult, in real life, to capture the stipulation that maximizing the expected monetary payoff is the sole—or at least dominant—objective. For most of us, that condition would indeed be captured if a $1,000,000 reward were at stake. But if we yield to practicality and substitute a nominal sum, then other goals—such as a desire to bolster the position one may be arguing for, or an inclination to experiment, or sheer whimsy—may prevail instead. In that case, it isn't enough to know you can correctly solve the problem of how to maximize your payoff, because your objective might diverge, in hard-to-predict ways, from the goal of maximal payoff.

Hence, without a fantastic simulator, Newcomb's Problem is not easy to enact in real life using a folk-psychological prediction. But doing so is at least possible in principle, and this possibility connects Newcomb's Problem to the Prisoner's Dilemma, as follows.

In (one version of) the Prisoner's Dilemma, two individuals face symmetric choices, and are assumed to be able to figure out the correct choice, and to have mutual knowledge of their mutual knowledge. Hofstadter (1985) terms such agents *superrational*—though, as the arithmetic example illustrates, superrationality with regard to an easy enough problem does not require unusual or implausible cognitive powers (contrary to what the word itself might suggest).

Each individual in the Prisoner's Dilemma faces a binary choice between cooperating with the other prisoner, or else defecting. As given at the start of the previous chapter, the prisoners' prospective jail sentences are: if both cooperate, five years each; if both defect, ten years each; otherwise, no penalty for the defector, but life imprisonment for the cooperator. Thus, given that the other cooperates, each does better to defect; given that the other defects, each also does better to defect. But both do better if both cooperate than if both defect.

By assumption, each of the two prisoners acts only for the sake of doing better personally, with no inherent regard for the other's welfare. And crucially, we stipulate that the choice to cooperate or defect has no other consequence of interest, beyond the immediate effect on the two prisoners'

respective sentences. (In contrast, the *iterated* Prisoner's Dilemma, discussed below, poses a series of trials in which a prisoner's choice in one trial affects the prisoner's reputation in the next, with possible consequences regarding others' subsequent cooperation with that prisoner.)

A paradox now arises as to your best course of action if you are one of the prisoners. On the one hand, you and the other prisoner each do better if cooperation is the right choice than if defection is the right choice. On the other hand, you yourself do better to defect than to cooperate, regardless of which choice the other prisoner makes. That is, given whichever choice the other prisoner actually makes, your cooperation causes you to do better than your defection causes you to do. But similarly, in Newcomb's Problem, given whatever amount of money is actually in the opaque box, your taking the transparent box as well causes you to fare better (by $1,000) than forfeiting it causes you to fare; yet, the previous chapter argued that taking the opaque box alone is nonetheless your more lucrative choice.

As Lewis (1979a), Horgan (1981), Leslie[3] (1991) and others have pointed out, Newcomb's Problem is structurally similar to the Prisoner's Dilemma (though Lewis advocates taking both boxes in Newcomb's Problem, and correspondingly recommends the uncooperative choice in the Prisoner's Dilemma). Superficially, the Prisoner's Dilemma differs from Newcomb's Problem in that the benefactor in Newcomb's Problem has no relevant goals of her own—the benefactor's postulated behavior, for whatever motivation, is to predict accurately, and to set up the large box accordingly.

3. Leslie (1991) proposes that *quasi causation* connects the behavior of two or more causally independent entities that operate according to "similar" causal factors, justifying the one-box choice in Newcomb's Problem, and cooperation in the Prisoner's Dilemma. But to do the work that Leslie requires of it, quasi causation must invoke a very broad, abstract sense of similarity—the simulator may be implemented in a different technology (transistors vs. neurons) and may use an entirely different algorithm than its subject, so that only their respective outputs (not the internal details that combine to produce the outputs) correspond sufficiently to let us construe one as an indication of the other. We must then explain why, for instance, if I will cross the street if and only if there is no dangerous traffic (as in secs. 5.5 and 5.6 above), my crossing is not abstractly similar to—and thus quasi-causative of—the absence of dangerous traffic, given the correlation that allows us to construe either as a reliable indication of the other in the specified situation. Without such an explanation, appeal to quasi causation may just amount to evidentialism by another name.

Still, from your standpoint as one of the choosers in the Prisoner's Dilemma, the other chooser's choice process, due to its postulated symmetry, is effectively a high-level simulation of your own choice process.

As in Newcomb's Problem, then, you face the prospect that the ''simulation'' of your choice—that is, the other player's actual choice—would very probably correspond to whichever your own choice is, despite the absence of a causal link. The simulation—the other's choice process—thus predicts your own choice (and vice versa), given the assumption of superrationality. And crucially, beyond a (merely evidential) prediction, there is a choice-supporting subjunctive link, a means–end link, from your choice to the other's corresponding choice, just as in Newcomb's Problem.[4] As in the folk-psychological version of Newcomb's Problem above, the means–end link composes two others:

- Given that you—faced now with a particular problem that lies within your competence—carefully solve the problem and act accordingly: if and only if you were to take a particular action, then that action would be the one regarded as correct by a competent choice process faced with that problem (assuming a unique correct action); and
- If and only if the action regarded as correct by a competent choice process were thus, then another competent chooser presented with the same problem would also regard that action as correct, and would act accordingly.

Each subjunctive link here applies to a range of different situations in which you (or another chooser) are faced with some problem that you are competent to solve. As in Newcomb's Problem, the first of these subjunctive links is acausal, the second causal (the second link summarizes the other competent chooser's machinery's response to the presented problem).

However, the acausal link here is not a past-predicate link of the sort that comes up in Newcomb's Problem with an atom-by-atom simulation. Hence I cannot help myself to the argument in section 5.6.4 to establish that the acausal link should be resistant to being explained away (the argument there was that a relative handful of causal regularities combine in exponentially many ways to account for all that occurs, and hence are especially

4. In contrast with the present account, Hofstadter (1985) defends cooperative action in the Prisoner's Dilemma only by appeal to an explicitly evidentialist criterion: what you do informs you of what others like you will do in like circumstances, so you should do what you want them to do.

widely applicable and explanatory; and substituting a coextensive past-predicate condition leaves a link just as applicable and explanatory). Without such an argument, we have no reason to be confident that the link here is in fact subjunctive (and hence a means–end link), rather than just evidential.

An analogous argument does apply here, though. Even for a single type of trivial problem, such as adding a pair of ten-digit numbers, competence to solve that problem involves a compact set of rules (expressing an addition algorithm) that apply to exponentially many situations (the number of *n*-digit numbers we might add is exponential as a function of *n*). More generally, the rules that express the algorithms for human-level intelligence apply to an even more vast set of situations that involve a vast number of problem types. Thus, as with basic causal laws, basic problem-solving rules apply to exponentially many situations, making those rules especially applicable and explanatory (as descriptions of the problem solver's behavior), and hence resistant to being explained away. Accordingly, we may expect that those links' presumptive status as means–end links remains standing according to the means–end-recognizing criteria proposed here. (As with causal and past-predicate links, though, this conclusion is not ironclad; I offer only a plausibility argument here.)

The subjunctive means–end link from one chooser's choice to the other's is seen most clearly here if we imagine that both choosers are implemented as computer programs, and in particular are identical computer programs, with identical inputs. Their respective outputs—their choices—must then be the same. If either chooser were to cooperate, so would the other; if either were to defect, so would the other. Neither chooser could make either choice with the reasonable expectation that the other's choice would differ.[5] And if we introduce a minute probability that the two computer programs diverge, then we minutely attenuate the link from one choice to the other corresponding choice, but the link is still there.

Even if the two computer programs run on different kinds of computers, with different kinds of processors and correspondingly very different behavior in terms of logic gates and so forth, the two still operate identically

5. Leslie (1991) makes the same point by imagining a universe consisting of two connected spatial halves, each a mirror image of the other. If you then face a Prisoner's Dilemma situation in which the other agent is your mirror image, the same strict correspondence holds as between two identical computer programs.

(or probably so, if we introduce slight random divergence) at a higher level of abstraction, a level that looks at the software that the hardware implements. And just as two different kinds of hardware can both run the same software, so too can two different computer programs behave identically at a still higher level of abstraction—as is the case when they both correctly solve some class of problems, even if by different algorithms. The same applies to people in lieu of computer programs, as with Hofstadter's superrationality, or even with a weaker parallel between the two choosers, such that both are *probably* competent to solve a class of problems including the present problem (and are both aware of their mutual competence and mutual awareness).

Thus, in the Prisoner's Dilemma as in Newcomb's Problem, there is arguably an acausal means–end link from your choice to your goal. Your cooperation is a means to the other's cooperation, because if defecting were the right choice for you, it would also be the right choice for the other, and similarly for cooperating. Yet given whichever choice the other makes, your cooperation *causes* you net harm. It is only the *other's* cooperation that causes benefit to you. The framework for justifying your own cooperation subjunctively, on the basis of what *would* then be the case concerning another's cooperation—rather than just on the basis of what your cooperation *causes*—is crucial to the ethical theory explored here.

The Prisoner's Dilemma is discussed in *game theory*, a field that studies desirable strategies for agents who interact in precisely circumscribed ways while pursuing their respective goals. But game-theoretic analyses of the Prisoner's Dilemma are tangential to the fundamental issue here. Before game theory is brought to bear, the problem must be formalized to designate the consequences of a player's potential moves. In the conventional formulation of the Prisoner's Dilemma, your choice and your opponent's are independent, and game theory trivially endorses defection as the dominant strategy (see, e.g., Binmore 1994). But if instead your choice were able to causally constrain the other player to make the same choice, the game would be formalized such that game theory trivially endorses cooperation (even if the causal influence is just probabilistic, but with sufficient probability).

The present claim is that (at least in the case of competent choosers) an acausal subjunctive link to the other player's choice likewise makes the other's choice a (probabilistic) consequence in the sense relevant to

means–end analysis, so the game formalization should be just as though the link were causal. Game theory per se does not address the means–end analysis; given that analysis, the game (so to speak) is over before game theory even makes its move.

It bears repeating that, as noted near the end of section 5.4, to speak of acausal means–end links is not to propose some sort of acausal ''force'' in the universe. Rather, the means–end relation—defined as the relation between action and goal such that the desirability of the goal rationally contributes motivation toward taking the action—is just an abstract relation among events sitting in spacetime; causality is another such relation, as is being correlated. The claim argued for here (and throughout chaps. 5 and 6) is just that the means–end relation is a subjunctive relation, which is broader than the causal relation between action and goal (but narrower than correlation).

7.2 Subjunctive Reciprocity

Although the Prisoner's Dilemma is traditionally cast as a story about two prisoners, many ordinary situations have the same structure. These are situations in which you stand to benefit from another's cooperation and vice versa, although you can do better for yourself by being uncooperative (as can your counterpart), provided that your uncooperativeness goes undetected and thus has no deleterious consequence for you (just as defection incurs no penalty in the Prisoner's Dilemma scenario, because the two prisoners choose independently, without knowledge of one another's choice, and so there is no relevant consequence apart from the effect on the prisoners' penalties).

Many authors have investigated how the Prisoner's Dilemma, extended to everyday situations of similar structure, might bear on the philosophy of ethics. Analyzing the Prisoner's Dilemma, we can explore whether it is rational, from the standpoint of self-interest, to behave altruistically, cooperatively, to act for others' benefit, given a situation in which doing so causes net harm to oneself, but in which everyone does better if everyone cooperates than if everyone doesn't. (Hofstadter 1985 offers one such discussion. Poundstone 1992 presents a survey of efforts dating at least as far back as 1950 using the Prisoner's Dilemma in its modern form, and to antiquity in other guises.) Horgan's application of Newcomb's Problem to the

Prisoner's Dilemma argues that there is a rational basis for behaving cooperatively in such situations, even presuming that one values only one's own welfare; promoting others' welfare becomes a subgoal of promoting one's own. Thus, there is a (subjunctive) means–end relation between the two, even in the absence of a causal link. We can say this not necessarily causal means–end relation in Prisoner's Dilemma situations constitutes *subjunctive reciprocity*.

7.2.1 Reciprocal Altruism Meets the Categorical Imperative

Subjunctive reciprocity suggests an intriguing prospect for the theoretical foundations of ethics. It offers a way to derive the rationality of acting altruistically, without requiring any specifically altruistic or ethical presupposition.[6] In this regard, the present theory of subjunctive reciprocity resembles the notion of *reciprocal altruism* as a basis for ethics. According to reciprocal altruism, one acts for the sake of one's own interests alone, but treating others well is rational because it influences others to do the same to oneself. Unlike reciprocal altruism, however, the present theory does not require one's altruistic behavior to *cause* reciprocity by others (although a causal link, if present too, can certainly contribute to the reciprocity). Rather, it suffices that one's behavior *subjunctively entails* reciprocity by other symmetrically situated competent choosers, as idealized in the Prisoner's Dilemma: if and only if cooperation were the right choice for you to make, it would be the right choice for them to make; and if and only if cooperation were the right choice for them to make, they would (more probably) do so.

Going beyond causality makes a crucial difference. The exclusive reliance on causal links renders reciprocal altruism an implausible basis for ethics, for the theory leaves open too many large loopholes. Reciprocal altruism only works to the extent that benevolent behavior is effectively rewarded, and predatory behavior punished. Punishments and rewards are doubtless an important factor in promoting benevolent behavior, but by themselves

6. Throughout this discussion, I use *altruism* in a technical sense to characterize any choice that the chooser takes because the choice benefits others, even though the chooser knows that the choice causes no net personal benefit, or even causes net personal harm. Thus, simply refraining from theft or murder—to the extent that one could profit from such an act and get away with it—is an example of altruism, in this sense.

they are inadequate, because cheating is often possible. If we all behaved in a predatory fashion to the extent that we could get away with it, to the extent that we could reasonably expect a net gain, never refraining from such profitable behavior simply because it is wrong or unfair, then (arguably) our level of cooperation would be dramatically less than it is in fact.

Purely causal self-interest theories are, metaphorically speaking, the hidden-variable theories of ethics. Just as so-called hidden-variable theories of quantum mechanics sought to resist the disintegration of objectivity by clinging to an inadequate technical basis—resulting in theories that were objective but false—so too with (causal) self-interest and ethics.

In real life, few if any proponents of reciprocal-altruism-based ethics seem to act on the permission to cheat that that doctrine, taken to its logical conclusion, implies (committing profitable murders, for example). Instead, they often rely on implausible exaggerations of the probability or degree of the rewards or punishments caused by good or bad behavior. When the principles become more abstract, however, as in the economic sphere, adherents of reciprocal altruism may take the logical conclusion of their position more seriously. Libertarian arguments for laissez-faire capitalism, for instance, are indeed dismissive of the pernicious consequences of some economic choices, caring only that those choices promote self-interest.[7]

To consider a complementary approach, Kant's doctrine of the *categorical imperative* manages to avoid the approval of cheating. The categorical imperative instructs each of us to act as we would (for our own sake) want others to act in comparable situations. I would not want others to victimize me, even if they can profit from it and get away with it. Therefore, the categorical imperative proscribes my victimizing others, even if I can profit from it and get away with it. But the categorical-imperative doctrine falters when it comes to explaining *why* one's desire about others' behavior rationally motivates behaving similarly oneself, in the absence of a causal link.

7. Utilitarian bases for capitalism—arguments that market forces promote the greatest good—are another matter, best suited for other books. For here, suffice it to note that even in theory, an unconstrained market does not promote the greatest good overall, but rather the greatest good weighted by the participants' relative wealth.

The present account of ethical foundations is a kind of synthesis of reciprocal altruism and the categorical imperative. Like reciprocal altruism, the present account grounds ethical concern in self-concern—behaving well toward others is a means to the goal of others' good behavior toward oneself. This grounding provides a crucial foundation: a noncircular justification for an ethical prescription. But although the form of the justification is similar to (but more general than) that of reciprocal altruism, the form of the prescription itself is that of the categorical imperative. Acausal means–end relations bridge the gap. If there is an inherent means–end relation between one's choice and a symmetric choice by another—even if one's choice does not *cause* any such reciprocity—then self-interest rationally motivates choosing as one wants others to choose—the categorical imperative—whether or not any causal link to a reward or punishment contributes to that motivation.

Of course, you may also have other sorts of reasons for treating others well, including simply caring about them. The relation between subjunctive reciprocity and other bases for cooperative behavior is discussed below in section 7.4.

7.2.2 Conditions for Subjunctive Reciprocity

How robust is the ethical prescription that derives from subjunctive reciprocity? Although the present theory plugs up some of reciprocal altruism's loopholes, do enough others remain to render the present theory inadequate as a foundation for ethics? Despite the arguably wide applicability of the Prisoner's Dilemma scenario, we need some extensions of that scenario to make the ramifications of subjunctive reciprocity more far reaching. Let us consider several such extensions in turn.

The Prisoner's Dilemma scenario requires that the other agent's choice not yet be known to you, except perhaps via foreknowledge of your own choice. Typically, though, we already know empirically approximately how benevolently the people we encounter tend to behave toward us. We reasonably expect those tendencies not to change drastically in most people in the near future (except perhaps as caused in reaction to our own unconcealed behavior). Thus, we already know, with reasonable reliability, approximately what the behavior of the symmetric choosers will be, at least

in the aggregate. The opaque-box Newcomb's Problem does not provide a reason to behave cooperatively in a Prisoner's Dilemma situation when we have empirically grounded foreknowledge of the likely degree of others' cooperation.

The transparent-boxes variant of Newcomb's Problem, however, provides a line of reasoning that meets the objection that the goal's achievement or nonachievement—here, the likely degree of others' cooperation—is already known. According to the conclusions drawn in the previous chapter, it makes just as much sense to act for the sake of a goal that you already know obtains (or already know does not obtain), provided that the same means–end relation exists between your action and your goal that would exist in the absence of that foreknowledge. You would like others to treat you well, even if they already know how others in turn treat them. Therefore, you should act correspondingly, behaving well even if you already know how others will treat you, because they would have no reason to do so if you were to have no reason to do so. The lesson extracted from the transparent-boxes variant thus provides a far more robust foundation for cooperation than does the original Newcomb's Problem—the transparent-boxes variant provides a foundation that is not undermined by already knowing the other persons' choices and thereby knowing the extent to which your goal has already been achieved.

One way to already know another's symmetric choice toward you is if that choice is in the past. It is commonplace, for instance, that if someone does you a favor, and you subsequently have an opportunity to perform the same favor for someone else, you may reflect on the favor you received and feel thereby motivated to act similarly. You may not explicitly regard the connection as a means–end relation (recall the discussion in secs. 5.2, 5.6.4, and subsequently about explicitly held theories versus built-in principles); nonetheless, in a practical sense, you act precisely as though recognizing the link as a means–end link. That is, the link connects your contemplated helpful action to an (already achieved) desired state—the state of receiving such a favor yourself—such that the desirability of that state provides incentive to take that action. The present account defends that practical means–end recognition: you respond as though there were a means–end link because there really *is* one, which your choice-machinery recognizes as such, regardless of your explicit opinion (if any) on the

question. (If there could not be a means–end relation when the goal lies in the past, then the account here could offer no way to justify an altruistic choice by someone who is just about to die.)

Furthermore, one may act well toward others but be treated badly oneself, or vice versa. And here too, one may already know of the others' choices of behavior toward oneself. Similar considerations arise in the dual-simulation transparent-boxes version of Newcomb's Problem, addressed in section 6.3. There, the correspondence that probably holds between your choice and the simulation outcome suffices to warrant taking just the large box (analogous to cooperating), even if the box is empty due to an unlikely simulation failure (analogous to being treated badly by others despite being cooperative yourself and despite a probable correspondence between your conduct and that of other similar choosers).

Again, of course, your choice does not *change* others' (causally independent) reciprocal choices from what they already are (or were, or will be). Instead, the claim is just that if you choose to act benevolently, we appropriately give that choice credit for any (actual or probable) instances in which others symmetrically choose to be benevolent to you; likewise, if you make the opposite choice, your choice appropriately gets credit for others' nonbenevolence. To give credit, in this sense, is just to acknowledge a means–end link: the choice is appropriately motivated by that which the choice gets credit for. (And of course, getting credit, in this sense, is not exclusive—that your choice gets credit for some outcome does not preclude giving credit for that outcome to other choices and events, including those that *cause* the outcome.)

The usual formulation of the Prisoner's Dilemma requires a degree of symmetry between two agents that applies only to a small subset of the situations that (intuitively) call for cooperative behavior. Both agents must be in a position to similarly affect one another, and to symmetrically contemplate the situation.

One relaxation of the symmetry requirement is immediately apparent. The situation need not pair symmetric agents A and B; instead, there might be an indefinite series of agents $\ldots A_i, A_j, A_k \ldots$, such that A_n is in a position to mistreat A_{n+1} for personal gain and get away with it. This scenario is not unrealistic. However powerful you may be, there will be (or there will have

been in the past—this, too, is relevant, as argued above) occasions when you are, symmetrically, at the mercy of others who stand to gain by acting nonbenevolently.

As with the two-agent formulation, the series of agents gives rise to a subjunctive link among agents' cooperation. To the extent that these others are the same sort of rational agent that you are—to the extent that you are all similar problem-solving mechanisms faced with similar problems—there is a rough isomorphism between your decision process and theirs. They have no more and no less reason to cooperate than you have in such situations. It is rational for you to behave well to those who are vulnerable to you, because you want to be treated well by those to whom you are vulnerable. Even if your treating others well in no way *causes* (possibly different) others to treat you well, it is still rational to do the former for the sake of the latter.

The superrationality requirement posited so far—that both agents (or the series of agents) are fully competent to solve the problem, and are aware of their mutual competence and of their mutual awareness—is a reasonable idealization in the Prisoner's Dilemma scenario. Real-life situations, though, presumably fall short of that idealization. In particular, if the present theory does turn out to be (at least roughly) correct, then competence to solve the problem would seem to require a knowledge and acceptance of (something like) that theory. But if only those who accept something like the present (arcane and controversial) theory were covered by the theory—if only those persons were entitled to be the beneficiaries of moral concern, or if they were the only available reciprocators—then the theory would be wildly implausible.

Fortunately, the requirement can be made less extreme. Agents who are probably more or less competent to solve the problem still fall within the scope of the above analysis—they need not have perfect mastery of the problem domain. The problem in question is to figure out whether cooperation is your rational choice (even in the absence of a causal link from your cooperation to the achievement of your goals) because of what (if that *were* your rational choice) a symmetrically situated agent would likewise recognize as rational. Any agent whose choice machinery recognizes a subjunctive means–end relation—leading to something like the influence of the categorical-imperative or golden-rule intuition—is thereby solving that

problem correctly (if the present theory is right), even if the agent's lack of a correct explicit theory leaves the solution somewhat vulnerable to seemingly sound counterarguments, and thus leaves the solution's influence somewhat tentative.

According to the present theory, we are agents whose choice machinery does indeed recognize subjunctive means–end links in Prisoner's Dilemma situations (albeit not with perfect reliability), leading to a categorical-imperative intuition, and a corresponding influence toward cooperation, regardless of what explicit theory, if any, we may subscribe to regarding means–end relations. We therefore can indeed solve the problem well enough for the mutual-competence analysis to apply.

As in the discussion above of almost-identical computer programs, if we lower the probability that a correct solution prevails for either of the Prisoner's Dilemma agents, we correspondingly weaken the means–end link from each agent's cooperation to the other's, but the link is still (partially) in force. Thus, among imperfectly rational beings, one may not be obliged to act for another's welfare to the same extent as for one's own—but there is still some degree of obligation. This conclusion accords, I think, with typical beliefs about such obligation. The more rational we are, the stronger the subjunctive-reciprocity link is, the stronger the reason to cooperate is, and the greater the resulting mutual benefit is. (Deficiencies in our rationality are partially compensated for by systems of causal incentives that augment the acausal component of reciprocity.)

The claim here is that rational moral regard reduces to subjunctive (not necessarily causal) reciprocity: roughly, you act as you want others to act toward you, because if that were the rational choice for you, it would likewise be the rational choice for them; and if they are (more or less) rational choosers, they would (probably) make what is the rational choice for them, which would then be to your advantage.

But if this analysis is right, you have incentive to "game the system," if possible, by choosing to follow a cleverly contrived policy that—if other choosers were to follow it too—would preferentially benefit you yourself, while minimizing your own deferral to others' interests. To what extent could that ploy succeed, under the current theory?

▪ For instance, it would be to your advantage if everyone were to abide by the policy *Act for the benefit of person X*, where *X* is defined as you in

particular (or is defined with respect to a set of properties that only you happen to exhibit). If you, as a more-or-less rational chooser, were to choose to act according to that policy, does it follow that that would be the rational choice, and that other rational choosers therefore would also act for the benefit of *X* (i.e., you)?

Fortunately, that inference does not follow. The subjunctive link between your choice and your potential reciprocator's choice depends on the (approximate) symmetry between your respective situations and between your respective choice processes. But being *X* does not have the same significance to *X* as to *X*'s potential reciprocator. If your rational choice were to act according to a policy rigged to benefit you preferentially, then the reciprocator's symmetric rational choice would not be to adopt that same policy of benefiting you, but rather to adopt a policy rigged to benefit the reciprocator. And that subjunctive consequence would not be to your advantage.

▪ But how about a policy *Act for the benefit of those who have property Y*, where *Y* is shared by you and by all your likely potential reciprocators? Then, *Y* may indeed play a symmetric role in those reciprocators' contemplations. For instance, when there is a dominant social class and a subordinate class, acting for the exclusive benefit of the dominant class is a policy that—if it were followed by potential reciprocators in the dominant class—would be beneficial to them. Property *Y* does not play a symmetric role among the subordinate class—but that doesn't matter to you as a *Y* if non*Y*s lack the power to affect you adversely.

Fortunately, there are problems with any such attempt to construe the set of beneficiaries narrowly so as exclude those who are vulnerable to you, but to whom you are not vulnerable:

— One problem is that if that were your rational strategy, it would also be the rational strategy of potential reciprocators who might benefit from a policy that uses an even narrower criterion *Z* that encompasses them and their reciprocators, but excludes you.

Inversely, if it were rational for you to *refrain* from choosing otherwise-arbitrary criteria designed to narrow the set of beneficiaries without excluding yourself, then it would be rational for your potential reciprocators likewise to refrain, which would be to your benefit, as it reduces the danger of their excluding *you*. This consideration rationally motivates using a broad criterion to say whose interests to respect, encompassing others who

may differ conspicuously from you—even others who do not act in (even implicit) recognition of subjunctive reciprocity.

On the other hand, the criterion in virtue of which to respect others can be delineated too broadly as well as too narrowly. A policy of respecting each entity's existence equally, for example, would have us fastidiously value a bacterium as much as we value a person. Following such a policy would not be a means to achieving your own goals, even (or perhaps especially) if we take account of any corresponding behavior by others that is subjunctively entailed by your following that policy.[8]

— Another consideration that militates against the arbitrary-narrowing strategy is that even if you are among a dominant group of *Y*s (such that you would benefit if you and others were to exclude non-*Y*s from being the beneficiaries of cooperation), your circumstances might have been otherwise. If various random events had transpired differently, you might have been a non-*Y* yourself.

The discussion of the dual-simulation transparent-boxes variant of Newcomb's Problem (sec. 6.3) argued that it can make sense to act in pursuit of an entire probability-distribution as to how things might have turned out—even if you already know how things *did* turn out—provided that the (sub)goal of your action cannot be made contingent on how things did turn out. (Here, the subgoal of the ploy under discussion is that it be the case that *Y*s, as competent choosers, act only for the benefit of *Y*s. The achievement of that subgoal cannot be made contingent on whether you yourself turned out to be a *Y*.)

The lesson of the dual-simulation transparent-boxes problem is thus consistent with the proposal of John Rawls (1999). Rawls advocates choosing a social policy as though under a *veil of ignorance* about your station—that is, you should choose a policy that you would want (for your sake) to be in

8. Additionally, as many authors have noted, even insofar as we are obliged to respect organisms' interests, most other species' interests differ qualitatively from our own. Organisms that are not prediction-value machines—or especially, those that are not even situation-action machines—lack any explicit interests at all, especially conscious interests. Many organisms, to be sure, do have interests in common with ours; aversions to physical pain and to stultifying confinement, for example, are commonplace among higher organisms. But ours is possibly the only species (or at least, one of few) to explicitly conceive of—and hence be able to explicitly desire—robust liberty, dignity, and even survival per se.

place if you were unaware of your actual circumstances, as though you were betting on the entire range of possible events that contributed to your present circumstances. The dual-simulation discussion offers an abstract decision-theoretic justification for betting on such a range of possibilities, regardless of which of those possibilities is already known to have come about.

Thus, under the present subjunctive-reciprocity account, we have reason to respect others' interests broadly, without restricting such respect to oneself or one's cohort. This reason is grounded in our own interests (appropriately including—as Rawls proposes—our interests as they might have stood under different circumstances). Yet we should not extend our respect so broadly (e.g., to bacteria) as to undermine all pursuit of our own interests. I make no pretense to have offered a detailed account of how to strike the appropriate balance; rather, my intent here is just to argue for an abstract foundation for such judgments, a way to ground them *at all*—by appeal to subjunctive reciprocity—without requiring unsupported ethical axioms that bear the ultimate weight of the entire ethical system.

Intuitively, there are two basic desiderata for a system of ethics:

- that the system prescribes behaving well toward other people, or prescribes behaving with respect for certain principles; and
- that the system provides a rationally compelling reason to behave as it prescribes.

Satisfying either of these is notoriously easy; satisfying both at once is notoriously hard. Nozick (1981) calls these desiderata the *pull* and *push*, respectively, of ethical systems, and points out that the main theoretical problem is to connect the two. Historically, humankind has often resorted to fantasies that bridge the gap: divine incentives (usually deferred to a supposed afterlife), karma (what goes around supposedly comes around), reincarnation (in a form that depends on your prior conduct), or an exaggeration of the extent to which one's conduct *causes* reciprocity.

A self-interest-based argument for behaving ethically—for respecting others, and (as discussed below) for respecting certain principles—is at once reassuring and disturbing. It is reassuring in that it does join the push with the pull. But it is disturbing because acting for pure self-interest

seems incompatible with acting ethically, even when the two happen to coincide. Kant, in particular, worried that appeal to self-interest renders respect for others merely contingent (on rewards and punishments being available, etc.), whereas genuinely ethical intention must be unqualified, categorical.

But consider the difference between *How would I like it if others treated me that way?* and *What's in it for me?* The first consideration sounds moral—it refers to a matter of fairness—but the second sounds invidiously calculating and selfish, and I think rightly so. We phrase things the second way when we think in terms of the punishments and rewards *caused* by our actions, the first way when we are thinking of acausally entailed consequences (which, however, we don't explicitly recognize as such, unless we happen to subscribe to the present theory). It is disturbing to think of ethics in terms of self-interest because our paradigmatic cases of acting in self-interest are examples of acting for the gain *caused* by our behavior. We properly reject *that* as an adequate foundation for ethics, recognizing that there are situations in which the ethically correct choice causes personal harm, not gain. And we properly take those situations to be important test cases for genuinely ethical behavior.

And indeed, the present account also rejects causal self-interest as the sole foundation of ethics. The apparent conflict between ethical motivation and self-interest arises, I suspect, from mistaking causal self-interest for self-interest generally. In paradigmatic examples of genuinely ethical conduct, even if there is no caused personal gain, the influence of *How would I like it if others...* questions is acknowledged, though not usually recognized as grounded in the pursuit of (subjunctively entailed) self-interest. That that influence turns out to be so grounded does not undermine its genuinely ethical nature; it should just change our theory of what genuinely ethical considerations are.

As argued above, we need not fear that this new sense of self-interest permits us to "cheat" the way purely causal considerations do. For here, the subjunctive entailment of reciprocity holds even in the absence of any causal link to the reciprocal benevolence. Reciprocity thus becomes inescapable, categorical—except, perhaps, in the case of a hypothetical entity so powerful as never to be, and never to have been, or to have had the possibility of being, very dependent, even indirectly, on others' benevolence.

But that much independence is so infeasible as to be contingent only by technicality.[9]

7.2.3 Consciousness and Subjunctive Reciprocity

Recall the question left pending at the end of chapter 2. Intuitively, we think of moral regard as being owed to something or someone if the entity is conscious and, as such, experiences desires, preferences, and so forth. But we can contrive a joke interpretation of the states within a rock such that the rock, according to that interpretation, has the sort of representations that constitute consciousness. Why doesn't the rock's consciousness (according to the joke scheme of sec. 2.3) entitle the rock to moral regard? Or conversely, why does your consciousness or mine, according to a non-joke interpretation of our states (that is, an interpretation scheme that passes Dennett's intentional-stance test) entitle us to moral regard?

Seeing ethics in terms of subjunctive reciprocity suggests an answer. The proposed basis for my deserving your moral regard is that your treating me well subjunctively entails others' doing likewise to you (which is to your benefit). But the same does not hold for a rock, regardless of the interpretation scheme we apply to it.

A joke interpretation of a rock, like the joke encryption of the lunch invitation in section 2.3, is very fragile—it applies to the rock as it is, but does not extrapolate to what the rock would be like under different circumstances. Although the rock's atoms' states for the past few minutes, accord-

9. Alternatively, an agent with only self-destructive goals, and a destructive inclination toward others, would have no rational basis in the current theory for treating others well, given that goal structure. (In the film *Dark Star*, space travelers use philosophical persuasion to try to defuse a "smart bomb." Alas, after due reflection, the bomb concludes that its sole purpose is to explode.) Our only recourse is to regard such a goal structure as pathological and to be thankful that it does not occur often (or more actively, to work to prevent it from occurring often). This move must be deployed with reluctance and caution, though—it is easy to claim hegemony for an ethical theory simply by proclaiming its rivals pathological. The danger is mitigated if the pathology designation is resorted to only in cases that are rare and extreme. And of course, when a destructive goal arises as a subgoal of other goals—whether tactically or strategically, as discussed in section 2.4.1—one can assess the (ir)rationality, as opposed to mere pathology, of its perceived facilitation of those other goals. Possibly, a self-destructive person's goal structure always turns out to be irrational in that sense.

ing to the joke interpretation scheme, map onto (say) the deliberations of a cognitive system similar to mine, it does not follow that if I were to make such-and-such choice in the Prisoner's Dilemma, the rock's atoms would then be in states that map onto *its* making the corresponding choice, under the joke scheme. The joke interpretation scheme thus does not hold up under counterfactual circumstances. The scheme does not support a subjunctive link, or (therefore) a means–end link, from my deliberations to the rock's ''deliberations.''

This consideration is not immediately decisive about the rock's deservedness of moral regard. As discussed in the previous section, you have reason to broadly designate the class of entities that you treat benevolently, in order to subjunctively entail a like strategy by others toward you. For that reason, you may act with some benevolence even toward entities that do not or cannot themselves behave benevolently, for the sake of subjunctively entailed reciprocation by others.

But as also discussed, it is not in your interests—even accounting for the subjunctively entailed behavior of others—to extend moral regard too broadly, to extend full respect even to each individual bacterium, for instance. And extending moral regard via arbitrary joke interpretations would be even more absurdly disadvantageous to you. Since any joke interpretation scheme is as legitimate as any other, a policy of cooperating with entities under joke interpretations would not permit you to arrive coherently at any preference of one action over another; any given choice could be either mandated or forbidden by the ramifications of some joke scheme or other. Therefore, subjunctive-reciprocity considerations—the foundation of ethical regard, according to the present account—do not argue for respecting a rock's interests as assigned to it by any joke interpretation scheme.

Thus we can, if we wish, acknowledge the rock's ''consciousness'' under a joke interpretation, but the construal is innocuous: it does not oblige us to respect the rock the way we should respect things that are conscious by a nonjoke interpretation. Alternatively—and, I think, more in keeping with common usage of the term *conscious*—we can construe deservedness of moral regard as part of the very definition of consciousness. In that case, even though the rock exhibits the requisite representations and self-representations (under the joke interpretation scheme), it still lacks a crucial quality that is partly definitive of consciousness. But either way—regardless of how we decide to define the term *conscious*—the rock, unlike

us, does not end up being entitled to moral regard (and thus, as always, substantive matters are not under the control of terminological decisions).

7.3 Ramifications beyond Altruism

Treating others nicely, as you would want to be treated yourself, is a central concern of ethics, but it is not the sole concern. Also within the domain of ethical thought are matters of retribution, and matters of principle—rights and responsibilities—that seem in some ways to transcend the considerations of altruism and reciprocity, causal or acausal. But an analysis similar to the argument for subjunctive reciprocity may also bear on these further considerations, as proposed, for example, by Hofstadter (1985), Hurley (1991), and Leslie (1991).

7.3.1 Reciprocity, Retribution, and Responsibility

Besides providing a means–end vindication of the golden-rule intuition, the present account offers a vindication of the less pleasant impulse toward retribution, often experienced as though its goal were somehow to retroactively prevent an already accomplished act of harm.

This retroactively oriented impulse is often dismissed as an irrational (albeit understandable) emotionalism—you can't, after all, undo a past wrong by retaliating. But by the present account, it is not necessarily futile to act for the sake of a past, known, and inalterable condition. Retribution may rationally have precisely the goal of not having undergone the harmful act in the first place (just as choosing just the large box in Newcomb's Problem—even if it is empty—may rationally have the goal of the box not having been empty in the first place, as discussed in sec. 6.3). Even though the offender had the opportunity to anticipate your retribution, that ''simulation'' of your chosen response failed to deter the offense. But acting for the sake of that failed retroactive consequence—that is, acting for the sake of the (albeit failed) deterrence—is still warranted, if the consequence had been probable enough.

Typically, of course, there is also an important purely causal consequence of, and motivation for, a retaliatory action, namely deterring *future* transgressions. But that consideration could not justify punishment in the absence of subsequent opportunities for wrongdoing (or if for some reason no potential future transgressor could learn that the punishment had been

carried out), whereas the current theory can.[10] There is a rational basis for retaliation even aside from any future consequences.

Indeed, retribution in retroactive pursuit of the failed goal can be rational even if the only available retribution is harmful to one's own interests as well as to the other party's. Game-theoretic doomsday scenarios (e.g., Shubik 1982) explore the paradoxical intuitions that come into play: self-harmful retaliation seems irrational in response to an already past attack (or, if your response in the event of an attack must be committed to in advance, but secretly and irrevocably, then it seems irrational to commit to self-harmful retaliation, since the opponent cannot be affected by your secret choice until it is too late for deterrence). Yet if self-harmful retaliation is not the rational choice, then there is no good reason for an attacker to fear such retaliation from a rational opponent, leaving that opponent without a credible deterrent. But this is just the by-now familiar conflict between a purely causal analysis of the consequences of an action, and a broader, subjunctive construal of means–end relations. If you, as a competent chooser, were to choose retaliation, then that would be what a competent choice process deems correct in such situations; and if a competent choice process were to deem that correct, then your opponent (if also sufficiently competent to discern correct choices) would so recognize, and would thus be deterred from attacking.

In real life, as just noted, the purpose of retaliation is largely causal: a transgression is punished in order to deter future transgressions, by causing a potential future transgressor to expect punishment as a likely consequence of a transgression. Similarly too with rewarding benevolent behavior in order to encourage future benevolence. But just as section 7.2 argued for the existence of a means–end link to reciprocal benevolence,

10. A distinction is often drawn between vengeance and justice, and properly so. The latter permits only those punishments that promote goals such as deterrence, the physical protection of others (e.g., by confining a violent predator), and rehabilitation. The present argument for the rationality of retribution says only that deterrence rationally includes the acausal component of deterring an already-past provocation. The emotional impulse to retribution may indeed reflect that acausal component, but the crude impulse is often excessive, leading, e.g., to indefinite cycles of reprisal. Nothing in the current defense of retribution denies the importance of subordinating revenge to a moderating framework of justice. And nothing in the current analysis supports the parity of an eye for an eye, especially if a lesser deterrent would be comparably effective.

even in the absence of a causal link (and even if the desired reciprocal benevolence is already past), so too with retribution in pursuit of deterrence.

It may be objected here that under the idealizing assumption of arbitrarily rational deliberation and conduct, the transgression to be punished would not have occurred in the first place, if the foregoing account is correct—because even apart from deterrence by anticipated retaliation, a rational enough agent should have been deterred by sheer altruism. But as we back off from presuming arbitrarily great rationality, substituting a more realistic model, we introduce the possibility that some means–end links are easier to recognize than others. In particular, a person who fails to recognize (what I claim to be) the acausal means–end link underpinning altruism may still be able to properly anticipate the likely retribution that a transgression would bring (insofar as transgressions are routinely observed to cause retaliation). Retribution or punishment can make sense in such situations.

Retribution thus bears on the notion of moral *responsibility*. By the present account: (1) we make choices, determinism notwithstanding; (2) it is rational to choose in a way that respects others' interests; and (3) it is rational to impose penalties in response to violations of the appropriate respect, in order to (causally or acausally) deter such violations. These factors constitute a sense in which we rationally hold people responsible for the ethical ramifications of their actions.

7.3.2 Cooperation When Each Individual's Influence Is Negligible

Suppose there are several agents whose cooperation is mutually beneficial, even though each stands to gain more by being uncooperative, given whatever choices are made by all the others. This state of affairs can obtain when the cost of cooperation is borne only by those who cooperate, whereas the benefits are distributed among all, whether they cooperate or not. As is well known, many common social situations are of this form. Hofstadter (1985), for instance, discussing cooperative decisions among rational agents who know of one another's rationality, lists examples like these:

- Reducing one's contribution to environmental pollution, even though the inconvenience to oneself outweighs the negligible benefit to oneself of the environmental improvement resulting from the reduction of one's own imperceptibly small contribution to pollution.

- Resisting an oppressive authority, even though those who do so may face retaliation that they could avoid by standing back and reaping the benefit of others' rebellion.
- Making the effort to vote, assuming (for the purposes of this example) that it is important for a particular candidate to get enough votes to win. Yet, in a large enough election, it is virtually impossible that your single vote would change who won—that would happen only if the vote were otherwise exactly tied. But *unless* your vote is a tiebreaker, that vote does not cause there to be a different victor than if you abstained.

The third example offers a further conclusion, beyond the prescription of benevolence argued for in the Prisoner's Dilemma analysis. The conclusion concerns the rationality of cooperating when the causal influence of each individual contribution is negligible, and their combined effect nonlinear.

Even if the foundation for regard for others is already established, a purely causal account of means–end relations falls short of providing a justification for voting in a sufficiently large election (presuming linear utility). If the importance of the election is presumed proportionate to the size of the electorate, then for large enough elections, expected-utility calculations cannot justify the effort of voting by appeal to the small but heavily weighted possibility that your vote will be a tiebreaker. The odds of that outcome decrease faster than linearly with the number of voters, so the expected value of your vote as a tiebreaker approaches zero—even taking account of the value to everyone combined, not just yourself. Given enough voters, then, the causal value (even to everyone) of your vote is overshadowed by the inconvenience to you of going out to vote.

Thus, even though the election's value per affected person remains constant under these assumptions as the electorate grows, a purely causal account of means–end connections cannot justify the effort of your voting in a large enough election, even presupposing an altruistic motivation that takes into account the benefit to *everyone* of the desired candidate's victory. Consequently, even presupposing a foundation for altruism, a purely causal account of means–end relations cannot justify certain important kinds of cooperation. But a generalization to subjunctive means–end relations suggests a way to provide that justification, by appeal to the acausal entailment of similarly situated others' behavior: if and only if you

were to have reason to vote, then so would they; and their combined votes, in turn, do have a decisive causal influence on the outcome.[11]

7.3.3 Reconciling Principle with Pragmatism

Ethical dilemmas often involve not only altruistic concerns, but also matters of principle that are not obviously reducible to pragmatic considerations.

For example, suppose you can tell a lie without getting caught, and the concrete consequences are beneficial to everyone. Let's say your acquaintance is hungry and eats only vegetarian food, but none is available, so you falsely state that the food you have is vegetarian. Your acquaintance would despise this deception, but (let us assume) has no way of finding out and is thus happy. But by the reasoning above, such a choice, even if successfully concealed, acausally entails similar behavior toward you in similar situations—behavior that you would despise and want to avoid. Moreover, such a choice would subjunctively entail its anticipation by others reasoning similarly, which would entail diminished trust in such assurances even when they happen to be true. Hence, the principle of not lying in such situations is well motivated, even when its application *causes* only inconvenience to all involved, in a particular situation.

The present approach suggests a solution to the apparent incommensurability of utilitarian concerns versus matters of principle. A fundamental controversy of ethics is whether an action that causes more good than harm can be inappropriate because it violates a compelling principle—or does the end always justify the means? In practice, virtually everyone seems to judge a large matter of principle to be more important than a small one of pragmatics, and vice versa—everyone except philosophers, that is. Kant, for example, found himself so committed to the priority of categorical imperatives over pragmatic concerns that he denied the right to tell a lie even in order to hide an innocent victim from a murderer (see Kant 1964, a translation of his 1785 treatise).

I maintain that the commonsense judgment, appealing to the magnitudes of the competing considerations, is correct. A coherent defense of

11. This discussion ignores the similar entailment of the *opposition's* voting behavior. We can sidestep that complication if we imagine a vote in which everyone is in agreement, but the vote will fail unless, say, 80 percent of the eligible votes are cast.

it has been elusive though, since matters of principle are seemingly so different from pragmatic concerns that it is difficult to justify a scale that includes them both—a scale that nonarbitrarily assigns matters of principle some finite, nonzero weight relative to pragmatic concerns. Among those who cherish coherent defenses, there is therefore a tendency to insist, like Kant, that one or the other kind of consideration must always prevail.[12]

But by the present approach, a ''matter of principle'' is just a matter of subjunctively entailed consequences. Principle thus becomes commensurable with pragmatics. As with utilitarianism, the task is to choose among alternative actions by tallying and comparing their respective consequences—but here, the consequences can be either causally or acausally entailed. Thus, this account suggests a vindication of principle-based ethics over pure (causal) utilitarianism, parallel to its vindication of altruism over (causal) reciprocal altruism:

- The account explains why, contrary to reciprocal altruism, it is sometimes rational to act for another's benefit, even when doing so causes only disadvantage to oneself.
- The account explains why, contrary to utilitarianism, it is sometimes rational to abide by some principle in a particular situation, even if doing so in that situation causes only disadvantage to *everyone*.

7.3.4 Self-Reciprocity

As Ainslie (2001) discusses, there are intrapersonal decisions that may pose the same sort of conflicts as those in Newcomb's Problem and Prisoner's Dilemma situations. Such conflicts occur, for example, when we *discount* the value of our distant-future well-being, preferring to achieve more proximal goals. Some emphasis on the near future is rational, in part because of the diminished reliability with which we can achieve more-deferred goals, other things being equal. But we regard as shortsighted or weak willed the excessive pursuit of immediate gratification in exchange for long-term detriment.

Ainslie cites evidence that the utility we (and other intelligent organisms) are hardwired to assign to future events decreases *hyperbolically* as a function of the time lag: a goal that is n days away is now assigned $1/c_1(n + c_2)$

12. An exception is Nozick (1993), who advocates using a weighted combination of different decision strategies.

the utility that the same goal would be assigned immediately beforehand, where c_1 and c_2 are (positive) constant scaling factors. So, for example, if it is now springtime, you might be tempted to sell your winter coat for just $5, since its deferred utility is diminished; come autumn, though, as cold weather looms, you might willingly repurchase the coat for $50. Often, though, we somehow resist the sort of temptation induced by hyperbolic discounting. How we do that is the subject of Ainslie's analysis.

Other discounting functions—for example, using the exponential discounting factor c^n preferred by most economists (where $0 < c < 1$)—might also have you sell your coat cheaply in advance. But exponential discounting at least maintains a certain consistency: if you would sell your coat now for $5, then you would also have preferred long ago that you will sell your coat now for $5 (and similarly, you would hope now to do so again next time around), because $ac^x > bc^y$ implies $ac^{x+k} > bc^{y+k}$. Not so for hyperbolic discounting: $a/c_1(x + c_2) > b/c_1(y + c_2)$ does not imply $a/c_1(x + k + c_2) > b/c_1(y + k + c_2)$. Here, a and b are the two competing utilities—for example, the utility of having $5 versus the utility of having a coat in the winter; x and y are the respective delays until those utilities are attained; c, c_1, and c_2 are constant factors as above; and k is the length of time in advance that the tradeoff is contemplated.

Thus, with hyperbolic discounting, you would wish, sufficiently far in advance of the potential transaction, that you will not sell your coat cheaply (because the anticipated benefit of the $5 is discounted almost as much as is the anticipated disadvantage of being coatless in the winter—because sufficiently far in advance, the latter delay is only slightly greater than the former, proportionately speaking, which is what matters to hyperbolic discounting). But when the potential transaction is imminent, hyperbolic discounting can switch that preference: the undiscounted immediate benefit of selling the coat exceeds the discounted anticipated harm of being coatless in the winter (or of having to repurchase the coat), leading you now to act contrary to how you wished in advance you would now act. How, then, might you resist the temptation of the shortsighted immediate gratification, and defer instead to your longer-term interests, given a hyperbolic discounting function?

Ainslie points out that because hyperbolic discounting can thus give you different preferences about the same future choice depending on how far in the future the choice is, reconciling this difference between your interests

at different times is effectively a negotiation among different versions of yourself, not unlike negotiations among distinct individuals with partially conflicting interests. But then what negotiating leverage does your future self wield against your current self? As Marvin Minsky has wryly asked, why should I care about my future self—what's he ever done for me?

This dilemma is by now familiar. A future version of yourself is indeed in no position to reciprocate your current version's benevolence. But the current version of you has the same reason to act for a future version as a past version had to act for the current version. To the extent that you are glad that you have acted previously to protect your current interests, even by some sacrifice at the time (or wish you had done so, if you did not), you have reason to act in that manner for your still-future interests. By the present account (which diverges from Ainslie at this point), there is actually a means–end link from your present foresighted action to the (current) fruits of your past (actual or now-wished-for) foresighted action—even though those fruits are already inalterably secured (or inalterably forfeited). In part, you rationally act now in the protection of your future interests—even beyond the (hyperbolically discounted) utility that your choice machinery directly assigns to those interests—for the sake of your (albeit already secured) *current* interests (just as you should act for the sake of the visible, already secured box content in the transparent-boxes problem of sec. 6.2).[13]

An illustrative instance of the conflict of present-versus-future interests arises in connection with repetitive addictive behaviors, as Ainslie points out. Because of hyperbolic discounting, you might value the immediate gratification of another cigarette or Big Mac more than you value the longer-term benefit of foregoing that reward—even though, if you could press a button right now that would somehow enforce an irrevocable decision to quit permanently, you would do so. This state of affairs is possible, Ainslie explains, because with regard to your *series* of future choices, you may strongly prefer the (albeit discounted) anticipated future benefits of quitting to the (almost as much in the future, hence almost as much discounted) anticipated sacrifice of future gratification. This preference may

13. Here again, of course, your choice machinery's recognition of this means–end link does not require your explicit agreement with the present theory. The recognition could occur even if your explicit beliefs reject the notion of a means–end link here.

even be strong enough (if it's a long enough series) to overcome too the undiscounted pleasure of an immediate fix, even though the discounted incremental long-term harm inflicted by the next fix itself does not prevail over the undiscounted immediate pleasure of that fix.

But in the absence of a magic push-button to enforce a decision about future behavior, this set of preferences motivates the familiar choice to do it just once more and then quit—a choice that, unfortunately and paradoxically, then repeats itself each next time, forestalling quitting forever. Still, each such choice follows rationally from a hyperbolically discounted preference scheme, together with a purely causal construal of means–end relations. (Or so I claim; Ainslie argues otherwise, as addressed below in sec. 7.3.5.)

The paradox resolves if, by the reasoning advocated here, you construe there to be an acausal means–end link from your current choice to your symmetric choices in the future (as well as to your choices in the past, as noted above; then, the pastward link lets you, e.g., resist the shortsighted temptation to sell your winter coat cheaply in the spring even if, say, you know you have only a year left to live and thus can't appeal to a long future series). By this account, you do, in effect, choose all at once (albeit with a possibility of exceptions, since means–end links can be probabilistic), as though by pressing a magic button. And as Ainslie documents (though without invoking acausal means–end links), thinking of choice consequences in such categorical terms has long been deemed a key to exercising the willpower to overcome temporary short-term preferences for the sake of longer-term goals.[14]

As mentioned above in section 6.2.4, one possibility (though not one championed by either Ainslie or me) is that there *is* a push-button of

14. Utility-based theories of choice face the question of why we would experience some choices as difficult to make, why we would sometimes vacillate or need "willpower." After all, tallying attenuated utilities (as given by schemas) is a matter of simple arithmetic. But an analytical choice machine—one that constructs schemas for itself—might engage in an ever-ongoing process of building schemas on the fly in the course of its deliberations—especially when existing schemas make assertions that are uncertain or contradictory. The contradiction between causal and partly acausal means–end intuitions in Ainslie-like hyperbolic-discounting scenarios might be a particularly strong example—as might a similar conflict of intuitions regarding moral reasoning. The processes involved in building schemas are not the focus of this book; but that is where one explanation of the difficulty of some choices might lie.

sorts to force future behavior to accord with a resolution you made. It might be implemented by what we can think of as a willpower module in your cognitive machinery. This hypothetical module can compete for control against the choice machinery (that is, against the apparatus that selects an action according to the desirability of what the machinery perceives would be the case if that action were taken). Once engaged (by your making a resolution), the module acts to defend the resolution against contrary later choices, physically overriding the subsequent choice process.

But even if such a module indeed exists, its influence is finite. Empirically, a preference for some action at the moment due to its anticipated consequence can often overwhelm the strength of an earlier resolution, and fortunately so—on the whole, it would probably be to our detriment if our prior decisions were irrevocable even when further reflection comes up with a better idea.

A resolution-enforcement module of limited influence might serve as an ad hoc crutch to compensate for our confusion about the means–end links involved in resolving disputes (between near and distant interests) that arise due to hyperbolic discounting. But in the example at hand, we still need a basis for deciding against the one-more-time choice if the influence of the hypothetical enforcement module does not happen to prevail. If we (correctly, I claim) construe there to be an acausal means–end link from the current choice to the other such choices, we obviate the need for a willpower module to (if it can) overcome the subsequent choice process. Instead, the subsequent choice process can just choose correctly. Thus, the present analysis depends on neither the existence nor the absence of a hypothetical willpower module.

Considerations of subjunctive self-reciprocity lead us once again to address Searle's view of consciousness (mentioned in sec. 2.3 above). Chapter 2 argued that what we observe when we observe our conscious cognitive processes turns out to be a collection of thoughts, perceptions, preferences, and so forth, recorded and played back by a kind of Cartesian Camcorder (under a nonjoke representation scheme, i.e., one that passes Dennett's intentional-stance test).

But incorporating Searle's point, the phenomenon we then observe is also correctly described as such a collection implemented by carbon-based

neurons (even if we do not introspectively observe what the implementation is, the implementation is in fact a property of the thing we do observe). If we think it substantive to ask which of those two correct descriptions truly corresponds to being conscious, then as discussed in section 2.3, we have likely fallen prey to semantic sleight of hand, smuggling in the implicit definition that being conscious entails (in part) being something whose interests should be respected by others (otherwise, there's no reason to balk even at a joke interpretation scheme that attributes consciousness to a rock).

Exposing the implicit definition, we find that the substantive question becomes: by virtue of what property—having a certain kind of representations, or having a certain kind of representations and a certain kind of implementation of them—should we value one another, or even our respective future selves? If one of your neurons—or all of them—were about to be replaced by functionally equivalent silicon prosthetics, would you have any less reason to care now about your postreplacement self than if your original neurons were to remain undisturbed?

By the present account, the reason to care is grounded in the subjunctive link between your choice and a symmetric choice made by another—or a symmetric choice made by yourself at a different time (you resist selling your winter coat for $5 in the spring, because you're glad you didn't similarly shortsightedly undermine your current interests by your previous choices—or because if you did, you wish you hadn't—and if you were to do so now, you would more likely have done so previously, or would more likely continue to do so in the future, even without a causal link from your action to those consequences). And that subjunctive link depends only on both of you using similar competence to solve symmetric problems. The specific composition (and other biochemical details) of the switching devices that implement the competence is as irrelevant to that subjunctive link as is your hair color (but see sec. 7.3.5 just below for a further contrast with Searle's view).

Therefore, even if deservedness of moral regard is a necessary condition for what we think of as consciousness (as discussed in sec. 7.2.3 above), the distinction between, say, being made of neurons versus transistors has no bearing on that condition, and thus no bearing on whether an entity is conscious.

7.3.5 Contrasts: Ainslie, Searle, and Kurzweil

I claim above that each next once-more-before-quitting choice would follow rationally from a hyberbolically discounted preference scheme, if not for a choice's acausal consequences. But Ainslie argues to the contrary that you can rationally overcome the one-more-time imperative, even given a hyperbolic discounting of exclusively causal consequences (the only kind of consequence he considers), and without appeal to a separate decision-enforcement mechanism (e.g., the magic push-button).

Instead, Ainslie argues, you can take account of the causal influence your current choice has on your future choices: if you decide now to abstain, you thereby inform your future selves of your propensity to abstain, which information helps influence your future selves to do likewise, rather than to consider the effort to quit futile, as they would be more influenced to do if instead you decide now (and on other such occasions) not to abstain, despite your long-term preference for quitting. So you can abstain now in part to motivate your future selves to do likewise, thereby helping to cause the long series of abstentions whose value (from your present perspective) sums to more (even after discounting) than the value of the undiscounted immediate gratification if you indulge now.

Ainslie offers an analogy to a multiplayer game in which all players are rewarded a little each time one player forfeits a somewhat larger reward, with indefinitely many such decisions made sequentially and publicly (so that for each player, the smaller rewards add up to more than the large reward, if enough other players forfeit their large rewards). Each player makes only one move. When your turn comes around, you have reason to forfeit the larger reward so that, by your precedent, you help influence the next player to expect the subsequent players to forfeit too—an expectation that is however partly contingent on the next player joining you in contributing to the evidence seen by the subsequent players. That contingency helps motivate the player to forfeit too (and likewise her successors). Your influence on the next player (and the ones thereafter) is the goal of your own forfeit.

In fact, though, if your various selves do base their decisions on purely causal links to hyperbolically discounted results—and if they all know they all do so—then you know your future selves will have no reason to be influenced by your present choice, because the precedent is not genuinely

informative: your next self would already be able to figure out (even with no recollection of the precedent) what the precedent was, just by recapitulating the previous self's deliberation process. Therefore, the current choice only causes the achievement or forfeit of the immediate gratification and the associated (deferred and therefore discounted) increment of longer-term harm—a trade-off that favors the gratification, under the present assumptions.

Thus, if you are sufficiently self-knowledgeable, your past choices to abstain could at best give your future selves evidence that (contrary to Ainslie's hypothesis, but consistent with the partly acausal means–end view) you were not basing your decisions on hyperbolically discounted, purely causal ramifications. But that falls short of providing what Ainslie's account would need (when considering a sufficiently self-knowledgeable individual): a reason to abstain if indeed your decisions *are* based on purely causal, hyperbolically discounted consequences. (Similar remarks apply to Ainslie's multiplayer analogy if the players are all competent to solve the problem correctly, and are mutually aware of their competence, as with Hofstadter's superrationality.)

To elaborate a bit on the rationale for Searle's (1980) position, his famous *Chinese Room* argument (in one variation) asks you to imagine you understand only English, but you perform the task of hand executing a computer program that's simulating the thoughts of a person who's having a fluent conversation in Chinese (it's possible in principle—although very cumbersome—for a person to take any computer program and figure out, step by step, what a computer would do if it were running that program, thereby executing the program by hand). Of course, the Chinese conversation would have to be trillions of times slower than in real time, but that's okay—this is only a thought experiment.

You're performing the computation, but you're not conscious of the meaning of any of the symbols or syllables you're manipulating. That is, the computation might emit the Chinese word for *dog* in an appropriate conversational context, but you, as you hand execute the computation, would have no consciousness of what the word means. Therefore, concludes Searle, such consciousness requires more than computation; a digital device performing the same computation as a conscious entity would not thereby be conscious, just as a computer simulation of digestion could not

digest an actual slice of pizza fed to the computer. *Some* artificial device, even using different machinery than a stomach, might well be able to digest pizza—but not just by performing a computational simulation of a stomach. Similarly for consciousness, according to Searle: the simulation lacks consciousness because it cannot understand the *meaning* of any of its computational manipulations, just as you don't understand Chinese as you hand execute the Chinese-speaking computer program.

As Ray Kurzweil and many others have replied (see, e.g., Richards 2002), your role in that computation is rather like that of, say, a neurotransmitter, which also has no consciousness or understanding of the computation it facilitates. That *you* don't understand Chinese does not establish that the computation you're laboriously executing doesn't itself understand Chinese, or isn't conscious of doing so. That, indeed, is the very question at hand.

In Richards 2002, Searle says this reply misses the point. The point is not just that the person hand executing the program doesn't understand Chinese, but rather that "the *reason* the man does not understand Chinese is that he does not have any way to get from the symbols, the syntax, to what the symbols mean, the semantics. But if the man cannot get the semantics from the syntax alone, neither can the whole computer" (p. 75; my emphasis).

However, that you do not get to the symbol's meaning does not imply that you cannot in principle. If you were to understand the complicated system you were hand executing, you would indeed understand Chinese, and far more—you would also understand how the program being executed understands Chinese. The program being executed is like a Rosetta stone on steroids. It would be replete with, for example, memories that encode visual and audio imagery of dogs, subjunctive representations of how dogs would be expected to act under various circumstances, taxonomies relating dogs to other organisms, and so forth—as well as representations of linguistic constructs related to dogs. So the dog-symbol's referent would in principle be deducible, at least if you had eons to devote to the reverse-engineering effort. The information is there, albeit not in a convenient, obvious format for you.

But the computer program, unlike its hapless hand simulator, need not laboriously decode the Rosetta stone. That's only necessary in order to translate the computation into someone *else's* understanding of its

symbols. Instead, the computer program already implements its *own* understanding of various symbols, and implements its own consciousness of that understanding via the recording and playback of its smart representations, as discussed in chapter 2.

Kurzweil, although a leading proponent of the possibility of machine consciousness, does not fully follow through, in my view, when it comes to the question of whether a nonbiochemical replication of a brain's computational processes would necessarily be conscious. He says in Richards 2002, "If there is one crucial insight that we can make regarding why the issue of consciousness is so contentious, it is the following: There exists no objective test that can absolutely determine its presence" (p. 45).

Kurzweil is speaking not just of the current state of the art, but of the existence in principle of such a test. We can establish by objective tests (performed by an external observer) that a given object *acts* just like a person. If we can probe under the hood in sufficient detail, we can even (in principle, though we don't know how to yet) establish in detail what computation the internal circuitry is performing. But we cannot, according to Kurzweil's remark, objectively establish what (if anything) its corresponding subjective experiences are like.

However, I believe the foregoing considerations outline precisely how you might confirm objectively that another entity (human or otherwise) has subjective, conscious experiences like yours. To summarize:

- We can in principle show that when you speak of (or otherwise report, or just think about) your conscious experiences, the events that your report ostensively points to—the events that in fact give rise to your report—turn out to be certain events in your brain (sec. 2.1).
- The events pointed to can be described at various levels of abstraction: in particular, as computational events (abstracting above the underlying implementation), or as biochemical events (not abstracting above the implementation). The computation includes the smart recording and playback of various events, using terms of representation that designate interrelatedness, implementing an *understanding* of those events, as discussed in section 2.2.
- You have reason to promote others' interests (or your own deferred interests)—even beyond the extent to which you may directly value those

interests—insofar as your benevolence subjunctively entails reciprocity by others (or by versions of yourself at different times).

- Participating in (partly acausal) subjunctive reciprocity is a matter of performing the right sort of computation—under a nonjoke interpretation scheme (sec. 2.3), but regardless of the particular implementation substrate for the computation.
- That an entity has the right kind of mental events, and in a way that obliges us to care about or respect its interests, is a central aspect of what we mean by *conscious*. Performing the right sort of computation (under a nonjoke scheme, but regardless of implementation) does qualify an entity to have its interests respected, fulfilling that aspect of the meaning of *conscious*.
- Performing the right sort of computation, in the foregoing sense, can in principle be objectively, externally verified (say by detailed, neuron-by-neuron monitoring of brain activity); hence, so can consciousness. Even today, without yet being able to monitor and understand the computation that implements consciousness, we can obtain persuasive (albeit not conclusive) circumstantial evidence just by something like a *Turing test* (Turing 1950)—an in-depth conversational interaction that indirectly probes the underlying thought process. External behavior similar to one's own, under a comprehensive enough set of circumstances, suggests (but does not prove) similar underlying computation (particularly, though not exclusively, if there is a shared biological origin). For proof, though, we would need to monitor that computation more directly, which we cannot yet do, but it is possible in principle.

When Kurzweil denies that an objective test can "absolutely" detect consciousness, he presumably means by contrast with the ordinary practical standard by which, say, the roundness of the earth can be objectively verified. On the contrary, I maintain that both the presence of consciousness and the shape of the earth can in principle be objectively established to an arbitrary degree of confidence.

7.4 But Can't We Simply Get Along? (Putting Reason in Its Place)

No doubt this chapter's account of ethical foundations will strike many readers as excessively ratiocinative. Surely our ethical behavior does not

rest on our painstaking derivation of complicated abstract arguments such as those presented here. A domestic animal may display the same sort of overt affection—or hostility—toward others as human beings do, but presumably without engaging in explicit philosophical deliberation. Quite plausibly, a person's altruism, or lack thereof, is determined primarily by the same combination of factors that (plausibly) determine a dog or cat's: a genetic predisposition to behave kindly (or nastily) to others under certain kinds of circumstances; and the circumstances one has been exposed to—particularly, how well one was treated during one's upbringing.

Moreover, we often feel that we *care* about others' well-being. We empathize with them. We do not feel that we just calculate how being nice to them promotes our self-interest, whether causally or (as discussed here) acausally. And finally, our behavior is shaped in large part by the transmission of social mores, both by the same sort of instruction and apprenticeship as transmits, say, knowledge of astronomy or botany, or the practice of cooking or carpentry; and also by a structure of social rewards and punishments, whether quite tangible (gifts and salaries; fines and imprisonment) or less so (social admiration or approbation, and the internalization thereof in feelings of pride or shame).

These factors have been explored at length, both theoretically and empirically. One important line of inquiry examines the evolution of altruistic behavior in an artificial setting, using the so-called *iterated* Prisoner's Dilemma. The iterated Prisoner's Dilemma has a series of trials, each like the original Prisoner's Dilemma, except that each participant remembers the opponent's choices from one trial to the next. The iteration allows a choice made in one trial to be punished or rewarded by an opponent's choices in subsequent trials. Axelrod (1984) has shown that genetic algorithms in iterated Prisoner's Dilemma situations can evolve cooperative strategies, suggesting a possible origin of biological entities' inclination toward self-sacrificing cooperation in some circumstances.

Frank (1998) has observed that human emotions, when they impel conduct seemingly contrary to rational self-interest, can often be seen as promoting rationally cooperative behavior. For instance, feelings of pride or shame can motivate a person to take personal risks on behalf of a community, when a safer course of action would be to sit back and share the benefits of someone else's initiative. Similarly, emotional attachment to one's offspring often inspires personal sacrifices on their behalf. Even an

impulse to avenge a friend or relative, jeopardizing one's own well-being in the process, might have evolved to promote the mutual protection that a credible deterrent brings. An innate tendency to such emotions might thus help implement an Axelrod-like evolved cooperative inclination. Social incentives can amplify (or attenuate) such inclinations. Pinker (2002) and Dennett (2003), for example, have elaborated these themes with a wealth of evidence and argument.

The influences of evolution and socialization, especially through the sculpting of emotions such as empathy, may seem to explain our moral behavior so comprehensively as to render ethical philosophy superfluous. What room, then, is left for the influence of an explicit theory of ethics on ethical behavior?

On closer inspection, the above factors, important though they are, explain much less than they appear to. First of all, the iterated Prisoner's Dilemma, unlike the original, uniterated version, creates situations where your cooperative behavior specifically *causes* later reciprocity, due to the effect your choices have on your opponent's cooperation with you on subsequent trials. As, for example, Dennett (1995) and Hofstadter (1985) discuss, the rationality of cooperation in the iterated Prisoner's Dilemma therefore does not explain why cooperation is rational in the absence of a causal path to a later reward for cooperating. That is, it does not explain why you should not victimize others when you are in a position to profit from doing so and not get caught.

Still, to the extent that evolution has rigged us with a disposition toward empathy and other cooperation-promoting emotions (as in Frank's account), we might simply behave cooperatively without needing a rationale for doing so (just as you do not need a rationale to keep your heart beating—it's simply built that way). But empathy is notoriously limited. We do not, for instance, grieve deeply each time we read of a stranger being murdered. And empirically, from the extent of violent, selfish, or predatory behavior in the world, we can see that whatever altruistic disposition our genes or upbringing may impose, it can in fact be overridden by other considerations, for better (violence used in self defense, perhaps, in small-scale or even large-scale conflicts) or for worse (harming people to rob them, or persecute them, or just for fun).

Moreover, there are many inclinations that, even if they result from specific genetic predispositions, we want to override. For instance, suppose

there is a genetic predisposition to alcoholism. If you learned that you had inherited the alcoholism genes, you would not necessarily resign yourself to becoming an addict, nor should you. A more sensible response would be to take special care to avoid the expression of that disposition. Or, a sense of empathy (whether hardwired or not) may disincline you to violate the bodily integrity of another. But if you are a surgeon, you must learn to suppress that aversion in order to make an incision through flesh. Dennett (1995), Gould (1981), and other critics of (some construals of) sociobiology point out that many putative genetically predisposed behavioral tendencies—for example, toward sexism or aggression in some situations—do not thereby constitute imperatives, either behavioral or ethical, even if the supposed genetic influences are real. But the same holds true for any genetic influences that tend to promote altruism or cooperation. The ethical imperative, if any, must still come from somewhere else.

In short, whatever emotional impulse we may have toward altruism and empathy, and to whatever extent it may be genetically hardwired, it does not obviate the need for explicit judgments about right and wrong. If it did not seem *correct* to act with kindness and fairness, even at a net personal cost—if there were no sensible reason for so acting, beyond a raw impulse to do so—then we would have reason to regard the raw impulse as pointlessly self-destructive—like a disposition to alcoholism or a purely visceral (so to speak) aversion to surgery—and we would have reason to attempt to overcome it. And it is plausible that that attempt would have at least partial success, since empirically an impulse to altruism or empathy can be and often is overridden, for reasons good and bad.

Thus, although a dog or cat is not in danger of having its friendly behavior diminished by a belief that the behavior lacks a rational foundation (because it presumably forms no opinion about rational foundations), humans may be subject to that risk. And conversely, a belief that our kindly inclinations are correct is likely to help cultivate and amplify those inclinations. An explicit belief in the obligation to treat others fairly enables us to go beyond what is compelled by the limited emotional experience of caring. Furthermore, we all experience temptations to do what is wrong if it profits us greatly. If there is an explicit belief that an obligation to be altruistic and principled is real, that it has a rational basis, then this belief presumably has some effect, at least in borderline cases. The belief is likely to push in one direction, whereas a belief that an altruistic inclina-

tion has no rational privilege over any other sort of inclination we might experience would likely push the other way.

It is not surprising that our built-in inclinations do not suffice to explain ethics. The biological evolution of altruistic behavior, construed as a learning process, can be viewed as an early step in reasoning about ethics—a step taken by evolution itself, rather than by an individual intelligence. But as with other learning carried out by evolution, we may expect this early step to be rudimentary compared to what we can reason about explicitly. By analogy, evolution has also implicitly learned about some basic properties of physical objects; this knowledge is embodied in whatever hardwired competence we have for perceiving, manipulating, and navigating among the objects in our ordinary environments. But however helpful a point of departure this hardwired knowledge may be, it is naive by comparison with the knowledge developed by physicists. It would be a terrible mistake to settle for our crude, hardwired version of either physics or ethics.

Similar considerations apply to socially inculcated tendencies toward cooperation. Many aspects of what we now recognize to be moral conduct began as revolutionary, unprecedented defiance of prevailing mores. For such progress to occur, social values themselves cannot be the ultimate origin of ethics. Consider the range of ethical beliefs and corresponding behaviors actually exhibited by large groups of people: from Nazism to humanism, from slavery and manifest destiny to freedom rides and Gandhian resistance. All these and more are demonstrably within the scope of human genetic, social, and psychological constraints. If a theory of ethics is to have finer resolution than this entire observed range, it must therefore appeal to more than social and biological constraints. It must invoke a sense of right and wrong that goes beyond a mere description of how our neural circuitry or social acclimation incline us to behave.

And we often do feel that our actions are grounded in part in an appeal to an abstract knowledge of right and wrong. Although you may dislike violence, you may nonetheless support, say, law enforcement, or a war or a revolution, due to being convinced of the justness of the cause. Or you may refrain from doing something that would benefit you—lying or stealing, for example—because you consider it wrong. Even if sufficiently strong self-interest overrides moral qualms—you may feel, roughly, that you were unable to resist the temptation to do it anyway—the moral qualms may still be felt to exert an influence, albeit not a decisive one. Explicit appeal

to principle is perhaps felt most strongly in the case of socially controversial matters—as democracy, slavery, executions, women's suffrage, and gay rights have been at various times, for example—when we are called upon to choose and defend a position among conflicting popular alternatives.

Of course, our introspection in such situations could be deceptive. It may be that our actions are caused by factors entirely other than beliefs about right and wrong, and that such beliefs merely occur to us as rationalizations of those actions. Quite plausibly this is often the case, just as more generally the reasons that we think are responsible for our doing or believing *anything* may just be retroactive rationalizations that substitute for the true cause. In many cases, though, when we see our beliefs or choices change under the weight of new evidence or arguments, we reasonably conclude that that evidence or argument likely caused the difference. Plausibly, then, explicit deliberations about right and wrong are at least sometimes influential in determining our actions.

Thus, at a minimum, explicit beliefs about right and wrong may exert influence when the balance among other factors is roughly even, or when one must take sides in a social conflict. More importantly, though, even if explicit ethical theorizing does not proximally influence our actions much in routine situations, the other factors that do operate in such situations may themselves be shaped in the long run by explicit ethical reasoning (among other factors). This consideration applies especially to social influences, punishments and rewards, and feelings of pride or shame. Even when we conform to social pressures without knowing their origin, we are acting under the extended influence of whatever reasoning (and whatever other factors) helped sculpt those pressures over the years and millennia. By analogy, our biological form is determined by the accumulation of our ancestral mutations, even though mutation rarely affects an individual reproductive step. Similarly, the culturally cumulative effect of explicit reasoning about ethics quite possibly predominates over other factors, even if the immediate impact of explicit reasoning is negligible at almost every step.

Attempts to logically derive ethical foundations without ethical presupposition should not be thought to suggest that such a derivation is necessary (or sufficient) to promote ethical conduct. Similarly, appeal to thought experiments involving agents with idealized rationality or idealized predic-

tive powers does not suggest that people would need to have such powers in order to behave ethically. And of course, we would be foolish to pretend that we humans *are* ideally rational and hence able to behave ethically by sheer exercise of reason. Alas, we must not forgo the systematic incentives and sanctions that, in reality, we need in order to supplement the influence of our limited rationality.

Still, I maintain it is both true and important that a sufficiently rational person would indeed have rational grounds, without prior ethical supposition, for benevolent and principled behavior, even if (unrealistically) all additional factors promoting such behavior were absent. It is important because if an arbitrarily rational person would find no reason for ethical behavior per se, that would be a reductio ad absurdum of the belief that one should behave ethically. Then, to the extent that we tried to base our actions on careful deliberation, we would be led away from ethical conduct, not toward it—benevolence and rationality would be adversarial rather than symbiotic.

It may well be easier to motivate our ethical conduct by appealing to intuitions such as *this is right*, *this is fair*, and *think about the other person's feelings*—rather than by the intellectual machismo of appealing only to abstract arguments about acausal means–end relations. Similarly, we would not need or want to try to motivate our every move on a bicycle by an analysis of Newtonian mechanics. Both in physics and in ethics, even if we accept the principles extracted from reasoning about idealized toy scenarios, the explicit application of those principles to everyday situations is often impractically complex. Anticlimactically, after all the analysis, we must revert to trusting our intuitions much of the time—intuitions that, I speculate, are implemented in part by means–end-recognizing machinery along the lines of what is sketched in chapter 5 above. (Dennett 1995 documents discussion of a similar point about intuition versus explicit reasoning at least as far back as the 1800s.)

Nonetheless, by understanding how our intuitions could possibly be competent to know the truth about physical objects, or about ethics—by knowing that there are underlying mechanical principles whose ramifications our brains could be computing, even if the details of the computations are not introspectively accessible—and by knowing the general form of those principles, we can better judge which of our intuitions to trust, and refine those intuitions. Knowing physics may not help much in riding

a bicycle, but it does help in designing a bicycle, not to mention a spaceship. And it helps us dismiss entire categories of spurious intuitions, such as those that pursue perpetual-motion machines or telekinesis. Knowing how our sense of balance works explains why we should trust it to stay upright while walking, but not while piloting an airplane inside clouds. Similarly, an account of ethical foundations can steer us away from grounding our choices in ancient mystical dictates, or in exclusive consideration of selfish causal consequences, while helping us understand why an intuitive balancing of categorical-imperative factors may be a more sound guide.

In sum, ethical theory, explicit belief about right and wrong, is not omnipotent in determining our behavior, but it is influential. Good theories of ethics can encourage us to behave well; bad theories can promote correspondingly unethical behavior. Grounding ethics in reciprocal altruism unduly encourages selfishness; ultimate reliance on social, legal, or religious tradition or authority tends to entrench the oppressive or persecutorial aspects of those institutions; and perhaps most insidiously, denial that there *is* a rational foundation for ethics exerts influence toward ethical relativism, which tends to imply that any adopted ethical standard is as good as any other—and thence toward ethical nihilism, the doctrine that there is no real distinction between right and wrong.

Both the one-box intuition in Newcomb's Problem (an intuition you can feel, e.g., in response to the play-money practice trials, even if you ultimately decide to take both boxes), and inclinations toward altruistic (or sometimes retributive) and principled behavior (inclinations you likewise can feel even if you end up behaving otherwise), involve what I have argued are acausal means–end relations. Although we do not (unless we accept the present theory) explicitly regard the links as means–end relations, as a practical matter we do tend to treat them exactly as only means–end relations should be treated: our recognition of the relation between the action and the goal influences us to take the action (even if contrary influences sometimes prevail).

I speculate that it is not coincidental that in practice, we treat these means–end relations as what they really are. Rather, I suspect that the practical recognition of means–end relations is fundamental to our cognitive machinery: it treats means–end relations (causal and acausal) as such because doing so is correct—that is, because natural selection favored machinery that correctly recognizes and acts on means–end relations without

insisting that they be causal, perhaps in the manner sketched in chapter 5. And that recognition is, I suspect, at least partly responsible for our golden-rule and categorical-imperative intuitions—regardless of whether we accept (or have even considered) an explicit theory of subjunctive reciprocity.

If we do not explicitly construe those moral intuitions as reflections of subjunctive means–end links, we tend instead to perceive the intuitions as recognitions of some otherwise-ungrounded inherent deservedness by others of being treated well (or, in the case of retribution, of being treated badly). This supposedly inherent deservedness (which would not be a purely mechanical property, but rather something extra) is, in my view, a false reification of that which inclines our underlying choice machinery to reciprocate (or retaliate)—that is, a false reification of the (subjunctively) instrumental value (to us) of our reciprocating (even when our reciprocating *causes* no benefit to us), an instrumental value that our choice machinery correctly recognizes, even if our explicit theory has no place for acausal links of that sort.

7.5 Summary

By the present account, the nature of the means–end relation turns out to bear on the perennial question of whether and how we can derive what *ought* to be from what *is*.[15] Even with respect to pursuing purely self-centered goals, deciding what action one ought to take for those goals requires, as a starting point, some built-in kernel of means–end recognition—a way to derive what *would* be from what is, in the choice-supporting sense of *would*, as discussed in chapters 5 and 6. But the reduction of the Prisoner's Dilemma to Newcomb's Problem (especially with transparent boxes) argues that an analytical choice machine—even with just self-centered goals and with just the built-in means–end principles that the machine needs to pursue those goals in mundane situations—could in principle—without any further, specifically altruistic supposition

15. Simply equating the two, by presuming that how things are is necessarily how they should be, is (rightly) known as the *naturalist fallacy* (it's natural so it must be right, says the fallacy). More generally, though, deriving *ought* from *is* refers to any line of reasoning from a collection of nonnormative (nonprescriptive, nonjudgmental) propositions to a normative conclusion—a conclusion as to what should be the case, or what one should do.

or inclination built in—derive *ought* from *is* in a way that prescribes cooperation with others, even when cooperation causes no personal benefit.

The key principle is that there is a subjunctive relation between your benevolence toward others, and others' toward you: if your competent choice process were to culminate in choosing to treat others well, then that would be the choice of competent choosers in such situations. And if that were the choice of competent choosers, it would tend to be the choice made, in particular, by choosers who are deciding how to treat *you*.

As in Newcomb's Problem, by the present account, this subjunctive link suffices to constitute a means–end link, even in the absence of a causal connection (that is, even if your treating others well cannot cause anyone to treat you well). This link provides a technical vindication of our golden-rule or categorical-imperative intuition that we should treat others as we ourselves would like to be treated—an intuition that, conceivably, is implemented in part by an underlying computation that correctly recognizes the subjunctive-reciprocity link as a means–end link. Such recognition would be an automatic consequence of what chapter 5 argued is a general mechanism that choice-making agents need if they are to recognize means–end links correctly, even in mundane situations.

Grounding ethical regard in subjunctive reciprocity also addresses a key question left hanging in chapter 2. It helps explain how it could be that purely mechanical entities such as ourselves, exhibiting the right sort of (conscious) computation according to the right sort of (intentional-stance) interpretation scheme, could have a rational basis to value one another, or even our respective selves (regardless of whether the computation happens to be implemented by neurons or by transistors or whatever). That explanation, together with the proposed resolution of the time-symmetry and quantum-mechanical paradoxes in chapters 3 and 4, and the account of deterministic choice in chapters 5 and 6, makes a case for a purely mechanical view of the universe (including ourselves)—a view that nonetheless makes sense of some of our key perceptions and intuitions about consciousness, time, choice, and ethics.

8 The Anticlimactic Meaning of Life

Toward the end of *Monty Python's The Meaning of Life*, Michael Palin briefly addresses the film's eponymous subject. "Here's the meaning of life. It's nothing very special. Try and be nice to people, avoid eating fat, read a good book every now and then, get some walking in, and try to live together in peace and harmony with people of all creeds and nations . . ."[1]

For all its counterintuitive aspects, the position set forth in this book is similarly anticlimactic in its conclusions about our lives. Our purpose for ourselves (the universe per se has no purpose for us) is to seek fulfillment of various sorts—emotional, intellectual, sensual—rationally guided in part by an obligation to respect and promote others' interests as well as our own—more or less as many people have believed for millennia. Our universe, for all its spectacular diversity and practical unpredictability, is ultimately orderly and reliable—mechanical, material, devoid of ghosts, magic, and miracles—as many scientists have suspected for centuries. The point of the present exercise has been less to propose new conclusions than to put forth partly new justifications for some longstanding doctrines.

Reconciling our purpose and value with our mechanical nature is a challenging task. A conventional, commonsense, contemporary popular view of reality goes something like this:

- The universe has a past, present, and future. Time flows forward, turning successive future points into the present, and successive present points into the past. The past is definite, fixed, inalterable. But the future is uncertain, free, variable—in part by virtue of our choices, which are not fully subject

1. "And finally," he adds, "here are some completely gratuitous pictures of penises to annoy the censors."

to any mechanical constraints, and in part by virtue of quantum indeterminacy, which may indeed help account for the freedom of our choices.

The (at least partly) nonmechanical nature of our consciousness is central not only to our ability to choose freely, but also to our personal value and our moral obligation to one another: a purely mechanical being cannot be aware, cannot feel, cannot choose, and can be neither the bearer nor the object of moral responsibility.

Apart from the foregoing, all events in the universe are subject to mathematical, mechanical physical laws.

This book defends an alternative perspective:

- All reality is grounded in mathematical, mechanical physical laws, which prescribe exceptionless regularities among physical events. Life and consciousness do not transcend physics, but rather are among its ramifications. There is no flow of time; rather, everything just sits statically in spacetime. Quantum phenomena involve the divergence and convergence of universe-states in configuration space, creating an interacting multiplicity of futures (and pasts). But the whole system still varies deterministically along the time axis. Quantum mechanics per se does not contribute to explaining consciousness, or vice versa.

Consciousness and choice are particular computational processes whose broad outlines are already comprehensible. Consciousness is a matter of having the right kind of smart representations and self-representations, under the right kind of interpretation scheme. Choice occurs when an agent selects actions for the sake of goals, in recognition of means–end relations. The most straightforward principles for discerning means–end links in mundane situations, if those principles are carried to their logical conclusion, sometimes prescribe acting for the sake of what your action cannot alter or even cause. In particular, such is the case regarding reciprocity in many golden-rule or categorical-imperative situations, leading to a rationale for ethical conduct—a rationale that is consistent with and follows from our particular mechanical nature.

Some of the details of this account are starkly counterintuitive—unavoidably so, as long as the pertinent intuitions are themselves mutually contradictory, creating paradoxes such as those explored in this book. Still, I believe that the conclusions here broadly accord with, and indeed vindi-

cate, our most fundamental scientifically informed intuitions about our nature and purpose—a desirable reality check for any philosophical stance that aspires to be taken seriously.

Here, then, are some concluding metaquestions—not about how we fit in with the rest of reality, but about how the entirety of reality fits in with even broader considerations.

8.1 Something for Nothing

Even granting all the foregoing, there remains a nagging question: why is there something rather than nothing? And more specifically, why is there this universe rather than some other? If the entirety of our universe is specifiable by a few simple equations describing a simple distinguished state and simple state-change laws, why this set of equations rather than some other? The positivists noted that there can be no evidence for our metaphysical status (because any process by which such status could provide evidence would render that status part of the world's physical state, hence would not be metaphysical), and invoked Wittgenstein's (1921) saying: ''Whereof one cannot speak, thereof one must be silent.'' This erudite way of saying *shut up* has a point, but needs some elaboration.

Perhaps, as some suspect, our universe's equations are the simplest that could give rise to beings like us who ask such questions (though this possibility is highly speculative). Although our equations would then be uniquely noteworthy, that still would not quite explain why those equations—rather than some others, or none—give rise to an actual universe. Perhaps, though, some (or even all) of those alternative universes are also real.[2] Or perhaps our universe just happens to be implemented as a simulation on someone's computer in some metauniverse.

We have no evidence of these alternatives, of course—indeed, we couldn't possibly. If our universe is exactly described by its own particular equations, then other universes (even one in which ours is implemented) can have no effect on us (because the equations leave no room for such an effect). Moreover, these alternatives would defer the question without ultimately answering it. Why do some (or all) such universes exist, rather than others (or none)? Or if there is a metauniverse in which ours is

2. See, e.g., Tegmark 1998.

implemented, why does *it* exist? (If there is an infinite sequence of such metauniverses, why does that infinite sequence exist, rather than something else, or nothing?)

At this point, it behooves us to step back from the question *Why is there something rather than nothing*, and to rearrange the punctuation a little: why, *is* there something rather than nothing? Is there really? Or, more to the point: what, if anything, would be the difference between there being something and there being nothing? What difference, if any, does the universe's existence make?

The question sounds odd. Of course it makes a difference—it makes literally all the difference in the world. We know the universe exists because we see and feel its myriad constituents. If the universe didn't exist, there would just be nothingness instead, and we wouldn't be wondering about it. It seems, then, that whatever equations describe our universe are somehow endowed with a "spark of existence" that gives substance to the otherwise vacuous equations.

But there is something familiar about the notion of a spark of existence that supposedly distinguishes our universe from a possible but nonexistent one. It is reminiscent of the dualists' spark of awareness (sec. 2.1)—the extraphysical essence that supposedly distinguishes a conscious being from one that (albeit externally indistinguishably) just goes emptily through the mechanical, computational motions. Both putative sparks face the same problem: even if they were real, we could not know of them, could not perceive them—because any such perception would constitute a miraculous violation of the definitive physics equations that already specify all our thoughts and perceptions; perceiving the extra spark would be responding to something beyond the equations themselves, if the spark itself is something beyond the equations themselves. Whatever it is that we perceive when we think we perceive the extraphysical or metaphysical spark, it cannot in fact be something extraphysical or metaphysical.

Or look at it the other way around: imagine an alternative set of equations that define an alternative, imaginary universe in which evolve intelligent, inquisitive beings like us. If we could compute what unfolds from those equations and watch what the eventually evolved beings say when they contemplate their world, presumably we would not then find them lamenting that their universe, for all its grandeur, unfortunately lacks that all-important spark of existence! On the contrary, of course, their universe

looks and feels to them as obtrusively, overwhelmingly real as ours does to us—and we would see them think so and say so.

Most importantly, they would think and say so for the same sort of reason as we do, a reason that must be rooted in the equations themselves (because the equations themselves ultimately specify every detail of those thoughts and words), without recourse to any spark of existence. And even if we did not carry out the computation of what the alternative equations specify—even if those equations were left out in the cold, unnoticed and unexamined—those equations would still be specifying a universe in which intelligent beings perceived and spoke of what they thought is a spark of existence, just as we do, and for the same reasons.

As with the gravity hypothesis in the mirror-asymmetry paradox back in section 1.2.3, it becomes superfluous to hypothesize a spark of existence, that is, some kind of grounding that distinguishes a real universe from an unrealized set of equations. It is superfluous because the ungrounded equations must already specify organisms who perceive their universe as real (i.e., who perceive the apparent spark), just as we do, and for the same reasons that we do. Those perceptions are already inherent in the equations themselves.

The reason for believing in an extra spark, I suspect, involves a simple confusion. If someone were to propose that there is a herd of invisible, intangible unicorns cavorting in our midst, unable ever to affect us (or to affect anything that affects anything … that affects us), our reaction would be that that absolute isolation, direct and indirect, from what we can interact with is precisely what distinguishes the imaginary from the real. An object's reality is thus relative to its manifestability to other real entities, starting with ourselves. Manifestability to imaginary entities doesn't count—the invisible unicorns are no more real by virtue of their intercourse with *one another*. Seemingly, then, something in the network of mutually interacting entities must be real for the interaction, or potential interaction, to confer reality on anything else in the network. Either it's all real, or it's all just imaginary.

In fact, though, anything that we can see, touch, and so forth, or perceive introspectively, strikes us as immediately real without our needing to preestablish that we ourselves are real. (Even a solipsist who draws the line at Descartes's *I think, therefore I am* at least grants the reality of her own introspectively perceived thoughts.) What we consider to be the universe

is just the transitive closure of such interactions: it's the things that affect our awareness, and the things that affect those things, and so on. But there is no need for us, or anything else in that network of interaction, to be endowed with a spark of existence, a kiss of being—nor, as just discussed, could we perceive such a spark even if it were there.

Thus, if by *real* we mean that which has a spark of existence that distinguishes it from imaginary alternatives, then there is no reason to think that any universe, including ours, is real. But if by *real* we mean that which can interact with us (or with things that interact with things that interact with us . . .), then of course the things in our universe are real (we can see them and touch them, directly or indirectly), and hence, so is their totality, the universe itself; whereas other universes and their contents are not real (although we can contemplate them, we cannot see and touch them, directly or indirectly). *Real*, in this sense, is an indexical term. From the standpoint of certain properties-of-equations that are *us*, the term *real* designates things that are fellow manifestations of the same equations, things that are here-in-these-equations-with-us.

Of course, to the denizens of another universe (specified by other equations), that universe is real (in the same indexical sense), rather than ours. But that indexicality does not imply that other universes are as real as ours—just as the indexicality of the phrase *here-in-this-room-with-you* does not imply that, say, my bicycle is here in this room with you as much as the book you are reading now is here in this room with you. Rather, the indexicality of *real* implies—exactly as common sense would have it—that only this universe, and the things it comprises, are real. That is, only those things are here-in-these-equations-with-us.

Insofar as something's reality involves the potential to interact with it (directly or indirectly), we can also speak of a matter of degree. Distinct quantum branches, for example, may have essentially no future interaction with us (though they are still part of our universe by virtue of, if nothing else, their past interaction with us), and are in that sense perhaps less real to us than things in our own branch. For that matter, though, most events halfway across the galaxy, or even halfway around the planet, are similarly less real to us than events in our immediate presence. Clearly, though, this watered-down sense of reality by degree, though it has some meaning, is quite distinct from the binary concept of just being real or not—that is, of being here-in-these-equations-with-us or not.

Even if there were (though we could have no evidence for it) something beyond what the equations specify—a spark of existence, a metauniverse that implements ours as a computer program, or whatever—it would be beside the point. For it could have nothing to do with any events in our universe (including the event of our deep conviction that our universe exists).[3] Distinguishing what is real from what isn't only makes sense *within* a universe—that is, within the network of things that physically affect one another. Similarly, the very concept of causal explanation makes sense only within such a network. Beyond that realm, as Wittgenstein remarked, we have nothing to say—indeed, nothing even to ask.

A century ago, we had little clue how matters of consciousness and ethics could be understood as aspects of a network of physical interactions. So it seemed at the time that abandoning metaphysics would sacrifice these matters too, leaving behind their mere shadows: behaviorism standing in for consciousness; convention, preference, or (what later became) sociobiology standing in for ethics. But Wittgenstein's call for silence does not correctly apply to the study of consciousness or ethics. On the contrary, there is much to say about those domains.

8.2 On Our Own

Conspicuously, many events around us are caused by willful agents' intentions—their desires and decisions. That the mechanical paradigm instead is universally applicable—and even underpins intentionality—could not have been conceived until the dawn of science. Not surprisingly, then, sentient deities—extrapolations of the familiar intentional paradigm—were our distant ancestors' explanation for the unknown, an explanation that remains popular even today. Thus, a brief closing word about religion is in order.

Big Judeo-Christian-Islamic ghosts manipulating the whole universe are as implausible as little Cartesian ghosts manipulating our pineal glands (sec. 2.1 above), for similar reasons. And there is no direct evidence of

3. Implementation by a computer program would be relevant to us only if the rest of the metauniverse sometimes intervened to divert the software from computing our universe's physical laws, introducing "miraculous" anomalies. But there is, to put it mildly, no evidence for any such complication. At least for now, it lies in the realm of fantasy.

God's existence, apart from some (mostly ancient) eyewitness reports that are less well confirmed than contemporary Elvis sightings. But a real god could easily provide real evidence. God might choose, say, to make the Earth disappear for an hour, while we all float safely in space. I for one, and presumably most other atheists—not to mention most theists—would take such an episode as a convincing demonstration that God is real. But instead, putative miracles are either trivial parlor tricks, or manifested in dreams or visions, or passed by word of mouth for decades or centuries before even being documented, or all of the above. They are thus the sorts of events that would surely be mistakenly testified to by someone, somewhere, whether they occurred or not. The testaments therefore amount to no evidence one way or the other.

The absence of compelling evidence for God is particularly perplexing from the standpoint of those religious traditions that hold that God did act in revelation to us (and at considerable sacrifice to boot). Why then did God not do a more convincing job of it? Some argue that direct proof would rob us of a free choice to accept or reject God. But on the contrary, a heroically dogmatic atheist could continue to reject God even in the face of overwhelming evidence, and could do so with no less warranted a leap of faith than is now required to be a theist. Moreover, merely acknowledging God's existence would not automatically compel allegiance or obedience, nor should it. Ample opportunity would remain for human choice about God, given real proof of God's existence. The choice would just become an informed one.

Lacking direct evidence for God, we must still consider more-abstract arguments. God's existence is thought by many to offer otherwise-elusive explanation and hope. The explanation is of the origin or purpose of the universe and life; the hope stems from God's putative love and moral authority.

But attributing the universe's existence to God has less than no explanatory force. Not only do all the same problems persist in a new form (Why is there this god rather than some other, or none? If the universe cannot simply have always existed, or have created itself spontaneously, how can God have always existed, or have self-created?), but new, irresolvable mysteries are piled on top (Just *how* did God create the universe?), and all gratuitously, since no evidence supports, nor explanation flows from, the postu-

lated mysterious entity and powers.[4] And as discussed in the previous section, we have no reason to believe that the universe has any underlying spark of existence; hence we have no need to postulate a god (or anything else) to explain such a spark.

Claims of God's moral authority are especially problematic. Except by the specious dictum that might makes right, why must an all-powerful creator's will constitute a moral imperative? Moreover, chronic divine acquiescence to the holocausts frequently visited upon the innocent is flatly and plainly irreconcilable with the possibility of a loving, omnipotent god, despite centuries of desperately contorted theological rationalizations that resemble the excuses of a battered spouse insisting the abuser is benign.[5]

Nevertheless, perhaps the most compelling reason to believe in God is the seeming absence of an alternative basis for ethics. For all its historical and scriptural bloodthirstiness,[6] theism does tend in part to endorse standards of compassion and responsibility that are crucial to genuine ethics. If there were no possible objective foundation for ethics apart from religion, the existence of God might indeed be less implausible than the ethical nihilism that would then be implied by atheism. As argued in chapter 7, however, those are not our only choices. Rather, ethical foundations may be derivable without theistic presuppositions.

Some fear the hubris of substituting our own reasoning for the supposed revelations of religion, especially with regard to matters of right and wrong. The danger is no doubt real, for there is much room for error. But the alternative peril is to mistake some of our worst impulses and prejudices (or those of our ancestors) for the perfect will of God, and in the deity's stark absence to appoint ourselves God's spokespersons and enforcers—hardly a

4. Bertrand Russell made such points in his famous essay "Why I Am Not a Christian" (1927).

5. The biblical Book of Job offers one explanation for the suffering of the innocent: God torments a noble, steadfast devotee in order to verify that the devotee will remain loyal no matter what harm God inflicts.

6. For example, in Numbers 31, God's followers obey the deity's command to attack the Midianites, demolish their cities, kill all their men, and capture all their women and children. At God's further instruction, the conquerors then slaughter all the male children, as well as all the female captives, except for the 32,000 virgins among them, whom God tells the attackers to keep for themselves. (So much for the moral cornerstone of Western civilization.)

more modest or cautious stance. With or without deities, then, let us just tread with due regard for our fallibility. Let us try to scrutinize our core beliefs with enough care and honesty and courage that when we are mistaken, we will have a decent chance to discover our error.

Finally, many religious individuals attest that their belief in God imparts an optimism that is otherwise beyond reach. This is a subjective matter, but for me the opposite holds. I can accept that we inhabit a world of both splendor and squalor, of comfort and brutality, and that we can work to improve the balance. But if I were convinced that a universe created by an all-powerful, all-loving deity could *still* be marred by recurrent agony and atrocity, then I would likely surrender in despair. Moreover, the notion that God is necessary for hope implies that life, back in godless reality, is hopeless. But it is not—it most emphatically is not—and I protest both the defeatism that says otherwise, and the escapism that denies the finality of physical reality, for better or worse.

If this brief dismissal of theism seems curt, I intend no insult. I deeply respect freedom of belief, including religious freedom. But religious beliefs deserve no exemption from the standards of criticism and debate that apply to other doctrines of social and intellectual importance. I respect religious beliefs and religious individuals, but I respectfully reserve the right to argue that they are mistaken.[7]

8.3 So Here We Are

Today, science is able to supplant creation mythology with ever-closer approximations to creation reality—a kind of nonfiction Genesis. Combining basic science with some of the arguments and speculations herein leads to the following summary.

The content of spacetime is static, unchanging. The entire universe can be expressed as a simple distinguished (''initial'') state, and a set of elegant, spatiotemporally local and symmetric basic physical laws according to which the state of a local piece of spacetime is a function of the nearby

7. My points of disagreement here do not apply to those religious persons who take God to be just a metaphor. There is no inherent problem with finding inspiration in great works of fiction, recognized as such, or in associated individual or communal rituals.

state of spacetime. By virtue of these basic physical regularities, local states interact.

Along the time axis, so-called later events are those further from the distinguished state than so-called earlier events. When macroscopic states that were previously uncorrelated interact, they thereby come to bear mutual information. This acquisition of mutual information establishes an apparent arrow of time pointing away from the distinguished state and toward increasing entropy, opposite the direction of self-reassembling broken eggs and anticipatory wakes.

The framework of the universe is quantum configuration space. Configuration-space branches diverge and converge. Interference upon convergence makes the superposed states in distinct branches detectable by statistical evidence within a given branch, even though only one element of such a superposition is observed in each branch.

Early on, at least in some configuration-space branches, matter coalesces, forming protons, atoms, stars, and galaxies. In places with the right balance of stability and perturbility—places such as portions of solar systems that harden into planets of moderate temperature—some interactions produce simple self-replicating patterns. Some replications are exact, others altered. Some altered copies are more robust than their predecessors—they or their successors last longer before annihilation, or propagate faster. More robust variations proliferate. This evolution yields teleological entities, life: their self-preserving properties are present *because* they are self-preserving—that is, because earlier entities with those properties were able to produce self-copies that also have those properties.

Some evolved variations tend toward elaborated machinery that achieves broader resilience. The very machinery of varied replication is subject to variation and elaboration, developing systematic, combinatorial self-representation, such as by the alphabet of DNA. Evolved variants also include organisms whose behavior is implemented by machinery that explicitly represents aspects of the organisms' behavior and environment. Such organisms are intelligent. Behaviors evolve that promote the survival of the organisms that exhibit those behaviors.

Among evolved intelligences, some can learn, modifying their behavior to better pursue the implicit goal of survival. Prediction-value schemes amplify the ability to learn, by making a variety of survival-promoting goals explicit, along with means to pursue those goals. Prediction-value

organisms make choices, selecting an action for the sake of what would be the case if that action were taken, assisted by a hardwired foundation for recognizing instances of that subjunctive means–end relation. An important special case of that relation consists of causal links; another consists of the subjunctive entailment of reciprocity. In advanced organisms, the recognition of subjunctive, partly acausal reciprocity supports golden-rule or categorical-imperative intuitions, a cornerstone of ethics.

Especially sophisticated prediction-value systems can synthesize new terms of representation and use them to express the state of the world and potential state transformations. They develop explicit memories and intricate communication, facilitating cultures that accumulate knowledge. Cultural amplification of intelligence accelerates technological invention, yielding spears, plows, paper, computers, and spaceships. Organisms develop science and philosophy, describing the physical world and the organisms' place in it, with respect to both their origin and their value, as in this very summary. In this very map, we are *here*.

Subsequently, the process of evolution can become even more elaborate, substituting intelligent design for natural selection, and substituting a new implementation substrate—transistors, say, instead of neurons—that is amenable to deliberate design, but poorly suited to natural selection's implicit teleology because simple electronic circuitry does not reproduce.

Seeing ourselves as a sliver of the physical world—as part of what happens to some distal, hardened fragments of the sun—is at first perplexing as to where our consciousness, our feelings, our ability to choose, and our moral value could sit, for physics per se is devoid of such things. But arguably, consciousness, emotion, choice, and ethical obligation turn out to be abstract properties of just the right kind of complex physical machinery. As such, they are indeed real, just as we'd thought all along.

Or to be more concise:

Lucid in the Sky
Inexorably,
the star convolves
until, recognizing itself,
it marvels thus.

References

Ainslie, G. 2001. *Breakdown of Will*. Cambridge University Press.

Albert, D. 2000. *Time and Chance*. Harvard University Press.

Allais, M., and O. Hagen., eds. 1979. *Expected-Utility Hypotheses and the Allais Paradox*. Reidel.

Antony, L., and N. Hornstein, eds. 2003. *Chomsky and His Critics*. Blackwell.

Aspect, A., J. Dalibard, and G. Roger. 1982. ''Experimental Test of Bell's Inequalities Using Time-Varying Analyzers.'' *Physical Review Letters* 49: 1804.

Axelrod, R. 1984. *The Evolution of Cooperation*. Basic Books.

Baars, B. 1988. *A Cognitive Theory of Consciousness*. Cambridge University Press.

Baum, E. 2004. *What Is Thought?* MIT Press.

Bell, J. 1964. ''On the Einstein–Podolsky–Rosen Paradox.'' *Physics* 1: 195.

Binmore, K. 1994. *Playing Fair: Game Theory and the Social Contract 1*. MIT Press.

Blackburn, S. 1998. *Ruling Passions: A Theory of Practical Reasoning*. Clarendon Press.

Boltzmann, L. 1966. ''On Zermelo's Paper 'On the Mechanical Explanation of Irreversible Processes.''' In S. Brush, ed., *Kinetic Theory, Vol. 2: Irreversible Processes*. Pergamon Press. First published in 1897.

Churchland, P. 1985. ''Reduction, Qualia, and the Direct Inspection of Brain States.'' *Journal of Philosophy* 82: 8–28.

———. 1990. ''Knowing Qualia: A Reply to Jackson.'' In P. Churchland, *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. MIT Press.

Clark, T. 2005. ''Killing the Observer.'' *Journal of Consciousness Studies* 12 (4/5): 38–59.

Cohen-Tannoudji, C., B. Diu, and F. Laloë. 1977. *Quantum Mechanics*. Wiley.

deBroglie, L. 1964. *The Current Interpretation of Wave Mechanics*. Elsevier.

Dennett, D. 1980. ''The Milk of Human Intentionality.'' *Behavioral and Brain Sciences* 3: 428–430.

———. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting*. MIT Press.

———. 1987. *The Intentional Stance*. MIT Press.

———. 1991. *Consciousness Explained*. Little, Brown.

———. 1995. *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. Simon and Schoster.

———. 2003. *Freedom Evolves*. Viking.

Deutsch, D. 1986. ''Three Connections Between Everett's Interpretation and Experiment.'' In R. Penrose and C. J. Isham, eds., *Quantum Concepts in Space and Time*. Clarendon Press.

———. 1997. *The Fabric of Reality*. Penguin Books.

———. 1999. ''Quantum Theory of Probability and Decisions.'' *Proceedings of the Royal Society of London* A455: 3129–3137.

DeWitt, B., and N. Graham, eds. 1973. *The Many-Worlds Interpretation of Quantum Mechanics*. Princeton University Press.

Drescher, G. 1991a. *Made-Up Minds: A Constructivist Approach to Artificial Intelligence*. MIT Press.

———. 1991b. ''Demystifying Quantum Mechanics: A Simple Universe with Quantum Uncertainty.'' *Complex Systems* 5: 207.

Eells, E. 1982. *Rational Decision and Causality*. Cambridge University Press.

Einstein, A., B. Podolsky, and N. Rosen. 1935. ''Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?'' *Physical Review* 47: 777.

Everett, H. 1957. '' 'Relative State' Formulation of Quantum Mechanics.'' *Reviews of Modern Physics* 29: 454.

Feynman, R. 1985. *QED: The Strange Theory of Light and Matter*. Princeton University Press.

Fine, K. 1975. ''Review of Lewis.'' *Mind* 84: 451–458.

Flavell, J., and E. Markman, eds. 1983. *Child Psychology: Volume 3, Cognitive Development*. Wiley.

Frank, R. 1998. *Passions Within Reason: The Strategic Role of the Emotions*. Norton.

Fredkin, E. 1990. ''Digital Mechanics.'' *Physica D* 45: 254–270.

Fredkin, E., and T. Toffoli. 1982. "Conservative Logic." *International Journal of Theoretical Physics* 21: 219.

Fritsche, J. 1999. *Historical Destiny and National Socialism in Heidegger's* Being and Time. University of California Press.

Gibbard, A., and W. Harper. 1977. "Counterfactuals and Two Kinds of Expected Utility." In C. A. Hooker et al., eds., *Foundations and Applications of Decision Theory*. Reidel.

Gleick, J. 1987. *Chaos: Making a New Science*. Viking.

Goodman, N. 1983. *Fact, Fiction, and Forecast*. 4th ed. Harvard University Press.

Gould, S. 1981. *The Mismeasure of Man*. Norton.

Greene, B. 2004. *The Fabric of the Cosmos: Space, Time, and the Texture of Reality*. Knopf.

Hájek, A. 2003. "What Conditional Probability Could Not Be." *Synthese* 137: 273–323.

Hayward, J., and F. Varela, eds. 1992. *Gentle Bridges: Conversations with the Dalai Lama on the Sciences of Mind*. Shambhala.

Hofstadter, D. 1979. *Gödel, Escher, Bach*. Basic Books.

———. 1985. *Metamagical Themas: Questing for the Essence of Mind and Pattern*. Basic Books.

Horgan, T. 1981. "Counterfactuals and Newcomb's Problem." *Journal of Philosophy* 78: 331.

Hurley, S. 1991. "Newcomb's Problem, Prisoners' Dilemma, and Collective Action." *Synthese* 86: 173.

Jackson, F. 1982. "Epiphenomenal Qualia." *Philosophical Quarterly* 32: 127–136.

Jeffrey, R. 1983. *The Logic of Decision*. 2d ed. University of Chicago Press.

Joyce, J. 1999. *The Foundations of Causal Decision Theory*. Cambridge University Press.

Kant, I. 1964. *Groundwork of the Metaphysic of Morals*. Translated and Analyzed by H. J. Paton. Harper and Row. First published in 1785.

Kavka, G. 1983. "The Toxin Puzzle." *Analysis* 43: 33–36.

Kent, A. 1990. "Against Many-Worlds Interpretations." *International Journal of Theoretical Physics* A5: 1764.

Leslie, J. 1991. "Ensuring Two Bird Deaths with One Throw: Quasi Causation and Newcomb's Problem." *Mind* 100: 73.

Lewis, D. 1973. *Counterfactuals*. Harvard University Press.

———. 1979a. ''Prisoners' Dilemma Is a Newcomb Problem.'' *Philosophy and Public Affairs* 8, 3: 235–240.

———. 1979b. ''Counterfactual Dependence and Time's Arrow.'' *Noûs* 13: 455–476.

MacKay, D. 1960. ''On the Logical Indeterminacy of a Free Choice.'' *Mind* 69: 28–41.

Metzinger, T. 2003. *Being No One: The Self-Model Theory of Subjectivity*. MIT Press.

Minsky, M. 1968. *Semantic Information Processing*. MIT Press.

———. 1986. *The Society of Mind*. Simon and Schuster.

Newell, A., and H. Simon. 1972. *Human Problem Solving*. Prentice-Hall.

Nietzsche, F. 1917. *Beyond Good and Evil*. Translated by H. Zimmern. Boni and Liveright. First published in 1886.

Nozick, R. 1969. ''Newcomb's Problem and Two Principles of Choice.'' In N. Rescher, Nicholas, et al., eds. *Essays in Honor of Carl G. Hempel*. Reidel.

———. 1981. *Philosophical Explanations*. Harvard University Press.

———. 1993. *The Nature of Rationality*. Princeton University Press.

Pearl, J. 2000. *Causality*. Cambridge University Press.

Penrose, R. 1989. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Penguin.

———. 1996. *Shadows of the Mind*. Oxford University Press.

Piaget, J. 1952. *The Origins of Intelligence in Children*. Translated by M. Cook. Norton.

Pinker, S. 1997. *How The Mind Works*. Norton.

———. 2002. *The Blank Slate*. Viking.

Poundstone, W. 1992. *Prisoner's Dilemma*. Doubleday.

Putnam, H. 1988. *Representation and Reality*. MIT Press.

Quine, W. V. O. 1969. ''Propositional Objects.'' In *Ontological Relativity*. Columbia University Press.

Rand, A. 1964. *The Virtue of Selfishness: A New Concept of Egoism*. New American Library.

Rawls, J. 1999. *A Theory of Justice*. Harvard University Press.

Richards, J., ed. 2002. *Are We Spiritual Machines? Ray Kurzweil versus the Critics of Strong A.I.* Discovery Institute.

Russell, B. 1903. ''Appendix B: The Doctrine of Types.'' In B. Russell, *Principles of Mathematics*. Cambridge University Press.

———. 1927. ''Why I Am Not a Chistian.'' In *Why I Am Not a Christian: And Other Essays on Religion and Related Subjects*. Touchstone, 1967.

———. 1945. *A History of Western Philosophy*. Simon and Schuster.

Searle, J. 1980. ''Minds, brains, and programs.'' *Behavioral and Brain Sciences* 1: 417–424.

Shubik, M. 1982. *Game Theory in the Social Sciences: Concepts and Solutions*. MIT Press.

Sokal, A. 1996. ''Transgressing the Boundaries: Toward a Transformative Hermeneutics of Quantum Gravity.'' *Social Text* 46/47: 217–252.

Tegmark, M. 1998. ''Is 'The Theory of Everything' Merely the Ultimate Ensemble Theory?'' *Annals of Physics* 270: 1–51.

Turing, A. 1950. ''Computing Machinery and Intelligence.'' *Mind* 59: 433–460.

Vandersypen, L. 2001. ''Experimental Realization of Shor's Quantum-Factoring Algorithm Using Nuclear Magnetic Resonance.'' *Nature* 414: 883.

von Neumann, J. 1955. *Mathematical Foundations of Quantum Mechanics*. Translated by R. T. Beyer. Princeton University Press.

Wallace, D. 2003a. ''Everett and Structure.'' *Studies in the History and Philosophy of Modern Physics* 34: 87–105.

———. 2003b. ''Quantum Probability and Decision Theory, Revisited.'' *Studies in the History and Philosophy of Modern Physics* 34: 415–439.

Wheeler, J. A. 1983. In *Quantum Theory and Measurement*, J. A. Wheeler and W. H. Zurek, eds., p. 182. Princeton University Press.

Wittgenstein, L. 1999. *Tractatus Logico Philosophicus*. Translated by C. Ogden. Routledge. First published in 1921.

Wolfram, S. 2002. *A New Kind Of Science*. Wolfram Media.

Index