Contents

Preface

I OFFER HERE no startlingly new views about the mind-body problem beyond what can be found in my earlier book *Mind in a Physical World* (MIT Press, 1998). Apart from some new material, on topics like substance dualism, the idea of reductive explanation, and the explanatory arguments for type physicalism, what the book does offer, I hope, is better focused and motivated arguments and a more clearly articulated overall view of the philosophical terrain involved. By and large I feel comfortable with the outcome, as it is presented here, of my toils and travails over the years; it is the kind of picture I feel I can live with, although there still are murky nooks and crannies that might harbor hidden difficulties and dangers.

As detailed below, all the chapters save chapter 2 originated as stand-alone lectures, and this meant that each had to be made largely intelligible on its own, with minimal references to outside sources. I have decided to preserve this character for each chapter. The six chapters of this book, therefore, are intended to be readable as independent essays as well as serve as links in the overall argument of the book. This, I believe, has both advantages and disadvantages. One advantage is that the reader can go over the book pretty much in any order he or she pleases, or pick the chapters that look interesting or promising. But there is also the disadvantage that, although I have tried to minimize this, there inevitably remains some

overlap of material from chapter to chapter (for example, similar material on reduction will be found in chapters 1, 4, and 6). I hope, though, that this does not obscure the overall structure of the book's arguments. (In this regard the reader might find the "Synopsis of the Arguments" helpful.)

The material presented in chapters 1, 3, 4, and 6 derives in part from a series of five lectures given as the Daewoo Lectures, in Seoul, Korea, in the fall of 2000, under the auspices of the Korea Academic Research Council with the support of the Chosun Ilbo. I am grateful to the director of KARC, Dr. Yong Joon Kim, for the invitation and to its staff for cordial and efficient support. I also want to thank the many dozen philosophers in Seoul who participated in the events as commentators, discussants, translators, and chairs.

Chapters 2 and 5 are based on the Taft Lectures delivered at the University of Cincinnati in the spring of 2003. I am grateful to Tom Polger, John Bickle, Bob Richardson, Don Gustafson, and other members of the U.C. philosophy department for their hospitality and stimulating discussion.

An early version of chapter 1 appeared as "Mental Causation and Consciousness: the Two Mind-Body Problems for the Physicalist" in *Physicalism and Its Discontents*, edited by Carl Gillett and Barry Loewer (Cambridge University Press, 2001; used here with permission). Chapter 2 was originally prepared as a reply to Ned Block's "Do Causal Powers Drain Away?" and appeared under the title "Blocking Causal Drainage and Other Maintenance Chores with Mental Causation" in *Philosophy and Phenomenological Research* in July, 2003 (used with permission). An earlier version of chapter 3, with the title "Lonely Souls: Causality and Substance Dualism", was included in *Soul, Body, and Survival*, edited by Kevin Corcoran (Cornell University Press, 2001). A version of chapter 6 was given at the 2002 Wittgenstein Symposium in Kirchberg, Austria, and appears in the conference proceedings for that year. A similar

version will appear as part of my contribution, "The Mind-Body Problem at Century's Turn," to *The Future of Philosophy*, edited by Brian Leiter, forthcoming from Oxford University Press.

I am indebted to Chris Hill, Terry Horgan, and Brian McLaughlin for valuable comments on chapter 5. My research assistant, Maura Geisser, has given me her usual dependable and efficient help. I am particularly grateful to Ian Malcolm, my editor at Princeton University Press, who has provided me with unfailingly friendly encouragement and capable support.

Synopsis of the Arguments

A STRONG PHYSICALIST outlook has shaped contemporary discussions of the mind-body problem. The aim of this book is to assess, after half a century of debate, just what kind of physicalism, or "how much" physicalism, we can lay claim to. My conclusion is that although we cannot have physicalism *tout court*, we can have something nearly as good.

Chapter 1 introduces the two principal challenges confronting contemporary physicalism. They are mental causation and consciousness. The problem of mental causation is to explain how mentality can have a causal role in a world that is fundamentally physical. The supervenience/exclusion argument shows that within a physicalist scheme, mental causation is possible only if mental phenomena are physically reducible. But is the mental reducible to the physical? In particular, can we give a reductive physicalist account of consciousness? This is the problem of consciousness. There are well-known, though by no means uncontested, reasons for thinking that phenomenal, or qualitative, consciousness cannot be physically reduced. In this way the two issues, mental causation and consciousness, become interlocked: the problem of mental causation is solvable only if mentality is physically reducible; however, phenomenal consciousness resists physical reduction, putting its causal efficacy in peril.

Chapter 2 presents a more detailed and improved formulation of the supervenience/exclusion argument, including an

explanation of its fundamental motivating idea. The argument is divided into two stages, each with a philosophical moral of its own, and I show that there are two materially different ways of completing the second stage. Two of the more important objections raised against the argument are discussed. The first concerns the overdetermination option. The proposal is that we accept any purported physical effect of a mental event as overdetermined by two sufficient causes, one mental and one physical. I reject this suggestion. The second objection claims that the supervenience argument proves too much—specifically that, if correct, it would show that either all causation drains down to the bottom level of microphysics, depriving all special sciences of causality, or, if there is no bottom microlevel, there can be no causation anywhere. This is the "causal drainage" argument. I offer a two-pronged reply. First, the supervenience argument, I remind the reader, has been designed as a reductio against antireductionism; its point is that antireductionism, in conjunction with certain plausible principles and propositions, entails mental epiphenomenalism, a conclusion most of us are strongly inclined to reject. If the drainage objection works, it only adds to the force of the reductio. Second, the drainage argument is shown to depend on some questionable assumptions. Thus, chapters 1 and 2 lay out mental causation and consciousness as the two central problems for the contemporary physicalist, and then motivate and strengthen the supervenience argument. This prepares the starting point of the overall dialectic of the book.

Ontological dualism positing immaterial minds has not been taken seriously in contemporary philosophy of mind. Possibly as a response to the difficulties posed by mental causation and consciousness, however, the dualist approach is now showing signs of a revival. In chapter 3, I take a backward look at the idea of minds as immaterial substances, to argue that the dualism of material bodies and immaterial minds is not a workable option for anyone. For this purpose I formulate a causal argument.

This argument shows that immaterial minds, if they existed, would be incapable of entering into any causal relations, whether with material things or with other immaterial minds. This makes them gratuitous posits with no explanatory purpose to serve. The discussion provides concrete content to the oft-expressed complaint that, on account of their "diverse" natures, it is difficult to conceive how immaterial minds and material things could causally affect each other. By eliminating immaterial substances, chapter 3 establishes ontological physicalism, the thesis that bits of matter and their aggregates exhaust the content of the world.

Immaterial minds having been banished, the main remaining question concerns the status of mental properties. In chapter 2, it was argued that if mental properties are to retain their causal efficacy, they must be reducible to physical properties. The saving of minds' causal efficacy is widely considered a presumptive (some would say, nonnegotiable) desideratum. However, it is not proper simply to assume the reality of mental causation as a premise and then derive from it the physical reducibility of the mental. The reducibility of mentality must be assessed on its own merit.

Some writers have claimed that although mind-body reduction cannot be carried through, this does not preclude the possibility of *reductive explanation* of the mental in terms of the physical/biological. In chapter 4, I examine reduction, reductive explanation, and their relationship, in general terms as well as in relation to the mind-body case. Reductive explanation of mentality apparently requires the derivation of psychological statements from statements about neural/physical states and processes. How can this be accomplished? There appear to be three presumptive possibilities: (i) via psychoneural correlation laws as auxiliary premises (Nagelian bridge-law reduction); (ii) via conceptual connections between mental properties and physical/behavioral properties (functional reduction); (iii) via a posteriori necessary psychoneural identities

as additional premises (identity reduction). It is easily shown that bridge-law reduction does not yield genuine reduction or reductive explanation, and I discuss how (ii) and (iii) might generate reductive explanations, or help close the "explanatory gap." Making use of (ii) requires the functional definability of mental properties in terms of physical/behavioral properties, and (iii) presupposes the availability of psychoneural identities. One side result of this discussion is that reduction and reductive explanation are more intimately tied to each other than sometimes supposed.

In chapter 5, I raise doubts about the availability of psychoneural identities by undermining a currently popular argument for psychoneural type identities, namely the explanatory argument. The claim is that these identities are warranted because of the indispensable role they play in generating explanations of phenomena that would otherwise remain unexplained. The argument comes in two forms. The first begins with the observation that psychoneural correlations are pervasively observed, and that they require explanations. It is then claimed that psychoneural identities (for example, "Pain = C-fiber stimulation") provide the "best" explanation of the correlations ("Pain occurs if and only if C-fibers are stimulated"), and therefore must be accepted. I show that this argument is seriously flawed. The second form of the explanatory argument forgoes the claim that psychoneural identities explain psychoneural correlations; on the contrary, the identities render the demand for explanations of the correlations incoherent and wrongheaded. Rather, we need these identities if we want to bring neurobiological theory to bear on the explanation of psychological facts. Pain causes distress. Why? Because pain is identical with neural state N_1, distress is identical with neural state N_2, and neurophysiology tells us a detailed story about how neural state N_1 causes neural state N_2. The general point is that psychoneural identities can generate neural/physical explanations of psychological regularities, and

that this is sufficient warrant for their acceptance. This argument, too, can be seen to be critically flawed. The fundamental problem with both forms of the explanatory arguments concerns the role of identities in explanations. I argue that in explanatory derivations the essential function of identities is to serve as "rewrite" rules (by putting "equals for equals"), and that they are not capable of generating explanatory connections on their own. Thus, both explanatory arguments fail. Moreover, type physicalism has yet to overcome more than a few familiar objections, such as the multiple realization argument and various well-known epistemic arguments. It is plausible to conclude that psychoneural type identities are not going to be available to underwrite an identity reduction of the mind, or to close the explanatory gap.

Chapter 6 begins with a recapitulation of the arguments of the previous chapters, with a view to determining the progress of the overall argument of the book up to this point. The position we have arrived at may be called *conditional physical reductionism*, the thesis that if mental properties are to be causally efficacious, they must be physically reducible. That is, to save mental causation we must reduce mentality. This is the challenge faced by physicalism. With reduction via psychoneural laws and via psychoneural identities having been ruled out, the only remaining reductive option is functional reduction. Considerations are offered in support of the view that cognitive/intentional properties, such as belief, desire, and perception, are functionally characterizable and hence reducible, but that qualia are not so reducible. (A position like this has been advocated by others as well.) According to conditional reductionism, therefore, the causal efficacy of cognitive/intentional states can be vindicated (this saves agency and cognition from epiphenomenalism), but epiphenomenalism still threatens qualia. The battle, however, is not entirely lost for qualia, for some crucial relational properties of qualia—in particular, their similarities and differences—are behaviorally manifest, making

their functional characterization possible. Moreover, it is qualia similarities and differences, not their intrinsic qualities, that make a difference to cognition and behavior. The intrinsic qualities of qualia cannot be captured within the physical domain, but that is no great loss. The final conclusion, therefore, is this: Physicalism is not the whole truth, but it is the truth near enough.

I

Mental Causation and Consciousness

OUR TWO MIND-BODY PROBLEMS

SCHOPENHAUER famously called the mind-body problem a "*Weltknoten*," or "world-knot," and he was surely right. The problem, however, is not really a single problem; it is a cluster of connected problems about the relationship between mind and matter. What these problems are depends on a broader framework of philosophical and scientific assumptions and presumptions within which the questions are posed and possible answers formulated. For the contemporary physicalist, there are two problems that truly make the mind-body problem a *Weltknoten*, an intractable and perhaps ultimately insoluble puzzle. They concern mental causation and consciousness. The problem of mental causation is to answer this question: How can the mind exert its causal powers in a world that is fundamentally physical? The problem of consciousness is to answer the following question: How can there be such a thing as consciousness in a physical world, a world consisting ultimately of nothing but bits of matter distributed over space-time behaving in accordance with physical law? As it turns out, the two problems are interconnected—the two knots are intertwined, and this makes it all the more difficult to unsnarl either of them.

Mental Causation and Consciousness

Devising an account of mental causation has been, for the past three decades, one of the main preoccupations of philosophers of mind who are committed to physicalism in one form or another. The problem of course is not new: as every student of western philosophy knows, Descartes, who arguably invented the mind-body problem, was forcefully confronted by his contemporaries on this issue.[1] But this does not mean that Descartes's problem is our problem. His problem, as his contemporaries saw it, was to show how his all-too-commonsensical thesis of mind-body interaction was tenable within an ontology of two radically diverse substances, minds and bodies. In his replies, Descartes hemmed and hawed, but in the end was unable to produce an effective response. (In a later chapter we will discuss in some detail the difficulties that mental causation presents to the substance dualist.) It is noteworthy that many of Descartes's peers chose to abandon mental causation rather than the dualism of two substances. Malebranche's occasionalism denies outright that mental causation ever takes place, and Spinoza's double-aspect theory seems to leave no room for genuine causal transactions between mind and matter. Leibniz is well known for having denied causal relations between individual substances altogether, arguing that an illusion of causality arises out of preestablished harmony among the monads. In retrospect, it is more than a little amazing to realize that Descartes was an exception rather than the rule, among the great Rationalists of his day, in defending mental causation as an integral element of his view of the mind. Perhaps most philosophers of this time were perfectly comfortable with the idea that God is the sole causal agent in the entire world, and,

1. For Gassendi's vigorous challenge to Descartes, see *The Philosophical Writings of Descartes*, vol. 2, ed. John Cottingham, Robert Stoothoff, and Dugald Murdoch (Cambridge: Cambridge University Press, 1985), p. 238.

with God monopolizing the world's causal power, the epiphenomenalism of human minds just was not something to worry about. In any case, it is interesting to note that mental causation is regarded with much greater seriousness by us today than it apparently was by most philosophers in Descartes' time.

In any case, substance dualism is not the source of our current worries about mental causation; substantival minds are no longer a live option for most of us. What is new and surprising about the current problem of mental causation is the fact that it has arisen out of the very heart of physicalism. This means that giving up the Cartesian conception of minds as immaterial substances in favor of a materialist ontology does not make the problem go away. On the contrary, our basic physicalist commitments, as I will argue, can be seen as the source of our current difficulties.

Let us first review some of the reasons for wanting to save mental causation—why it is important to us that mental causation is real. First and foremost, the possibility of human agency, and hence our moral practice, evidently requires that our mental states have causal effects in the physical world. In voluntary actions our beliefs and desires, or intentions and decisions, must somehow cause our limbs to move in appropriate ways, thereby causing the objects around us to be rearranged. That is how we manage to navigate around the objects in our surroundings, find food and shelter, build bridges and cities, and destroy the rain forests. Second, the possibility of human knowledge presupposes the reality of mental causation: perception, our sole window on the world, requires the causation of perceptual experiences and beliefs by objects and events around us. Reasoning, by which we acquire new knowledge and belief from the existing fund of what we already know or believe, involves the causation of new belief by old belief. Memory is a causal process involving experiences, physical storage of the information contained therein, and its retrieval. If you take away perception, memory, and reasoning, you pretty much take away

all of human knowledge. Even more broadly, there seem to be compelling reasons for thinking that our capacity to think about and refer to things and phenomena of the world—that is, our capacity for intentionality and speech—depends on our being, or having been, in appropriate cognitive relations with things outside us, and that these cognitive relations essentially involve causal relations. To move on, it seems plain that the possibility of psychology as a science capable of generating law-based explanations of human behavior depends on the reality of mental causation: mental phenomena must be capable of functioning as indispensable links in causal chains leading to physical behavior, like movements of the limbs and vibrations of the vocal cord. A science that invokes mental phenomena in its explanations is presumptively committed to their causal efficacy; if a phenomenon is to have an explanatory role, its presence or absence must make a difference—a *causal* difference. Determinism threatens human agency and skepticism puts human knowledge in peril. The stakes are higher with mental causation, for this problem threatens to take away both agency and cognition.

Let us now briefly turn to consciousness, an aspect of mentality that was oddly absent from both philosophy and scientific psychology for much of the century that has just passed. As everyone knows, consciousness has returned as a major problematic in both philosophy and science, and the last two decades has seen a phenomenal growth and proliferation of research programs and publications on consciousness, not to mention symposia and conferences all over the world.

For most of us, there is no need to belabor the centrality of consciousness to our conception of ourselves as creatures with minds. But I want to point to the ambivalent, almost paradoxical, attitude that philosophers have displayed toward consciousness. As just noted, consciousness had been virtually banished from the philosophical and scientific scene for much of the last century, and consciousness-bashing still goes on in some quarters, with some reputable philosophers arguing that

phenomenal consciousness, or "qualia," is a fiction of bad philosophy.[2] And there are philosophers and psychologists who, while they recognize phenomenal consciousness as something real, do not believe that a complete science of human behavior, including cognitive psychology and neuroscience, has a place for consciousness, or that there is a need to invoke consciousness in an explanatory/predictive theory of cognition and behavior. Although consciousness research is thriving, much of cognitive science seems still in the grip of what may be called methodological epiphenomenalism.

Contrast this lowly status of consciousness in science and metaphysics with its lofty standing in moral philosophy and value theory. When philosophers discuss the nature of the intrinsic good, or what is worthy of our desire and volition for its own sake, the most prominently mentioned candidates are things like pleasure, absence of pain, enjoyment, and happiness—states that are either states of conscious experience or states that presuppose a capacity for conscious experience. Our attitude toward sentient creatures, with a capacity for pain and pleasure, is crucially different in moral terms from our attitude toward insentient objects. To most of us, a fulfilling life, a life worth living, is one that is rich and full in qualitative consciousness. We would regard a life as impoverished and not fully satisfying if it never included experiences of things like the smell of the sea in a cool morning breeze, the lambent play of sunlight on brilliant autumn foliage, the fragrance of a field of lavender in bloom, and the vibrant, layered soundscape projected by a string quartet. Conversely, a life filled with intense

2. A frequently cited source of consciousness eliminativism is Daniel C. Dennett, "Quining Qualia," in *Consciousness in Contemporary Science*, ed. A. J. Marcel and E. Bisiach (Oxford: Clarendon, 1988). See also Georges Rey, "A Question about Consciousness," in *Perspectives on Mind*, ed. Herbert Otto and James Tuedio (Norwell, MA: Kluwer, 1988). Both are reprinted in *The Nature of Consciousness*, ed. Ned Block, Owen Flanagan, and Güven Güzeldere (Cambridge, MA: MIT Press, 1997).

chronic pains, paralyzing fears and anxieties, an unremitting sense of despair and hopelessness, or a constant monotone depression would strike us as terrible and intolerable, and perhaps not even worth living. In his speech accepting the Nobel Prize in 1904, Ivan Pavlov, whose experiments on animal behavior conditioning probably gave a critical impetus to the behaviorist movement, had this to say: "In point of fact, only one thing in life is of actual interest for us—our psychical experience."[3] It is an ironic fact that the felt qualities of conscious experience, perhaps the only things that ultimately matter to us, are often relegated in the rest of philosophy to the status of "secondary qualities," in the shadowy zone between the real and the unreal, or even jettisoned outright as artifacts of confused minds.

What then is the philosophical problem of consciousness? In *The Principles of Psychology*, published in 1890, William James wrote:

> According to the assumptions of this book, thoughts accompany the brain's workings, and those thoughts are cognitive of realities. The whole relation is one which we can only write down empirically, confessing that no glimmer of explanation of it is yet in sight. That brains should give rise to a knowing consciousness at all, this is the one mystery which returns, no matter of what sort the consciousness and of what sort the knowledge may be. Sensations, aware of mere qualities, involve the mystery as much as thoughts, aware of complex systems, involve it.[4]

In this passage, James is recognizing, first of all, that thoughts and sensations, that is, various modes of mentality and consciousness, arise out of neural processes in the brain. But we can only make a list of, or "write down empirically" as he says, the observed de facto correlations that connect thoughts and

3. Ivan Pavlov, *Experimental Psychology and Other Essays* (New York: Philosophical Library, 1957), p. 148.

4. *The Principles of Psychology* (Cambridge, MA: Harvard University Press, 1981), p. 647; first published in 1890.

sensations to types of neural processes. Making a running list of psychoneural correlations does not come anywhere near gaining an explanatory insight into why there are such correlations; according to James, "no glimmer of explanation" is "yet in sight" as to why these particular correlations hold, or why indeed the brain should give rise to thoughts and consciousness at all.

Why does pain arise when the C-fibers are activated (according to philosophers' fictional neurophysiology), and not under another neural condition? Why doesn't the sensation of itch or tickle arise from C-fiber activation? Why should any conscious experience arise when C-fibers fire? Why should there be something like consciousness in a world that is ultimately nothing but bits of matter scattered over spacetime regions? These questions are precisely the explanatory/predictive challenges posed by the classic emergentists, like Samuel Alexander, C. Lloyd Morgan, and C. D. Broad—challenges that they despaired of meeting.

These, then, are the problems of mental causation and consciousness. Each of them poses a fundamental challenge to the physicalist worldview. How can the mind exercise its causal powers in a causally closed physical world? Why is there, and how can there be, such a thing as the mind, or consciousness, in a physical world? We will see that these two problems, mental causation and consciousness, are intertwined, and that, in a sense, they make each other insoluble.

I now want to set out in some detail how the problem of mental causation arises within a physicalist setting.

The Supervenience/Exclusion Argument

Mind-body supervenience can usefully be thought of as defining *minimal physicalism*—that is, it is a shared minimum commitment of all positions that are properly called physicalist, though it may not be all that physicalism requires. As is well

known, there are many different ways of formulating a supervenience thesis.[5] For present purposes we will not need an elaborate statement of exactly what mind-body supervenience amounts to. It will suffice to understand it as the claim that what happens in our mental life is wholly dependent on, and determined by, what happens with our bodily processes. In this sense, mind-body supervenience is a commitment of all forms of reductionist physicalism (or type physicalism), such as the classic Smart-Feigl mind-brain identity thesis.[6] Moreover, it is also a commitment of functionalism about mentality, arguably still the orthodoxy on the mind-body problem. Functionalism views mental properties as defined in terms of their causal roles in behavioral and physical contexts, and it is evidently committed to the thesis that systems that are alike in intrinsic physical properties must be alike in respect of their mental or psychological character. The reason is simple: we expect identically constituted physical systems to be causally indistinguishable in all physical and behavioral contexts. It is noteworthy that emergentism, too, appears to be committed to supervenience: If two systems are wholly alike physically, we should expect the same mental properties to emerge, or fail to emerge, in each; physically indiscernible systems cannot differ in respect of their emergent properties. Supervenience of emergents in this sense was explicitly noted and endorsed by C. D. Broad.[7]

5. See Brian McLaughlin, "Varieties of Supervenience," in *Supervenience: New Essays*, ed. Elias Savellos and Ümit Yalçin (Cambridge: Cambridge University Press, 1995).

6. Herbert Feigl, "The 'Mental' and the 'Physical'," in *Minnesota Studies in the Philosophy of Science*, vol. 2 (Minneapolis: University of Minnesota Press, 1958); J.J.C. Smart, "Sensations and Brain Processes," *Philosophical Review* 68 (1959): 141–56.

7. C. D. Broad, *The Mind and Its Place in Nature* (London: Routledge and Kegan Paul, 1925), p. 64. For more details on why supervenience must be an ingredient of emergence, see my "Being Realistic about Emergence," in *The Emergence of Emergence*, ed. Paul Davies and Philip Clayton (forthcoming).

Mind-body supervenience has been embraced by some philosophers as an attractive option because it has seemed to them a possible way of protecting the autonomy of the mental domain without lapsing back into antiphysicalist dualism. Just as normative/moral properties are thought to supervene on descriptive/nonmoral properties without being reducible to them, the psychological character of a creature may supervene on and yet remain distinct and autonomous from its physical nature. In many ways, this is an appealing picture: while acknowledging the primacy and priority of the physical domain, it highlights the distinctiveness of creatures with mentality—creatures with consciousness, purposiveness, and rationality. It reaffirms our commonsense belief in our own specialness as beings endowed with intelligent and creative capacities of the kind unseen in the rest of nature. Further, this view provides the burgeoning science of psychology and cognition with a philosophical rationale as an autonomous science in its own right: it investigates these irreducible psychological properties, functions, and capacities, discovering laws and regularities governing them and generating law-based explanations and predictions. It is a science with its own proper domain untouched by other sciences, especially those at the lower levels, like biology, chemistry, and physics.

This seductive picture, however, turns out to be a piece of wishful thinking, when we consider the problem of mental causation—how it is possible, on such a picture, for mentality to have causal powers, powers to influence the course of natural events. Several principles, all of which seem unexceptionable, especially for the physicalist, conspire to make trouble for mental causation. The first of these is the principle that the physical world constitutes a causally closed domain. For our purposes we may state it as follows:

> *The causal closure of the physical domain.* If a physical event has a cause at t, then it has a physical cause at t.

There is also an explanatory analogue of this principle (but we will make no explicit use of it here): If a physical event has a causal explanation (in terms of an event occurring at *t*), it has a physical causal explanation (in terms of a physical event at *t*).[8] According to this principle, physics is causally and explanatorily *self-sufficient*: there is no need to go outside the physical domain to find a cause, or a causal explanation, of a physical event. It is plain that physical causal closure is entirely consistent with mind-body dualism and does not beg the question against dualism as such; it does not say that physical events and entities are all that there are in this world, or that physical causation is all the causation that there is. As far as physical causal closure goes, there may well be entities and events outside the physical domain, and causal relations might hold between these nonphysical items. There could even be sciences that investigate these nonphysical things and events. Physical causal closure, therefore, does not rule out mind-body dualism—in fact, not even substance dualism; for all it cares, there might be immaterial souls outside the spacetime physical world. If there were such things, the only constraint that the closure principle lays down is that they not causally meddle with physical events—that is, there can be no causal influences injected into the physical domain from outside. Descartes's interactionist dualism, therefore, is precluded by physical causal closure; however, Leibniz's doctrine of preestablished harmony and mind-body parallelism, like Spinoza's double-aspect theory,[9] are perfectly consistent with it. Notice that neither the mental nor the biological domain is causally closed; there are mental

8. The closure principle should be distinguished from the thesis of physical determinism to the effect that every physical event has a physical cause. Physical causal closure should make sense even if some physical events don't have causes.

9. Here I am referring to the bare mind-body ontologies associated with Leibniz and Spinoza; I rather doubt that Leibniz's metaphysics of monads or Spinoza's metaphysics with God as the only substance would allow real causal relations even within the physical domain.

and biological events whose causes are not themselves mental or biological events. A trauma to the head can cause the loss of consciousness and exposure to intense radiation can cause cells to mutate.

Moreover, physical causal closure does not by itself exclude nonphysical causes, or causal explanations, of physical events. As we will see, however, such causes and explanations could be ruled out when an exclusion principle like the following is adopted:

> *Principle of causal exclusion.* If an event *e* has a sufficient cause *c* at *t*, no event at *t* distinct from *c* can be a cause of *e* (unless this is a genuine case of causal overdetermination).

There is also a companion principle regarding causal explanation, that is, the principle of explanatory exclusion, but we will not need it for present purposes. Note that the exclusion principle as stated is a general metaphysical principle and does not refer specifically to mental or physical causes; in particular, it does not favor physical causes over mental causes. It is entirely neutral as between the mental and the physical. For our purposes, it will be convenient to have on hand a generalized version of the exclusion principle.

> *Principle of determinative/generative exclusion.* If the occurrence of an event *e*, or an instantiation of a property P, is determined/generated by an event *c*—causally or otherwise—then *e*'s occurrence is not determined/generated by any event wholly distinct from or independent of *c*—unless this is a genuine case of overdetermination.[10]

The second principle broadens causation, or causal determination, to generation/determination simpliciter, whether causal or of another kind. The intuitive idea is the idea of an event or

10. In chapter 2 this broader principle will be dispensed with in formulating the supervenience argument.

state, or a property instantiation, owing its existence to another event or state—or, to put another way, the idea that one thing is generated out of, or derives its existence from, another. What I have in mind is very close to the fundamental notion of causation, or determination, that I believe Elizabeth Anscombe was after in her *Causality and Determination*.[11] Causation as generation, or effective production and determination, is in many ways a stronger relation than mere counterfactual dependence,[12] and it is causation in this sense that is fundamentally involved in the problem of mental causation. Another way in which a state, or property instance, is generated is supervenience; the aesthetic properties of a work of art are generated in the sense I have in mind by its physical properties. So are moral properties of acts and persons generated by their nonmoral, descriptive properties. It is the relation that sanctions the assertion that something has a certain property *because*, or *in virtue of* the fact that, it has certain other properties that generate it. I have argued elsewhere for the causal/explanatory exclusion principle;[13] I believe that the fundamental rationale for the broader principle is essentially the same, and that anyone who finds the former plausible should find the latter equally plausible.

It is quick and easy to see how these principles create troubles for mental causation for anyone who accepts mind-body

11. Cambridge: Cambridge University Press, 1971. Reprinted in *Causation*, ed. Ernest Sosa and Michael Tooley (Oxford: Oxford University Press, 1993).

12. It is in some respects weaker than counterfactual dependence; in cases of preemption and overdetermination, generative causation may hold without counterfactual dependence. The two notions are not strictly comparable, and that is why the counterfactual accounts of causation continue to have difficulties with preemption and overdetermination, showing, in my opinion, that our core idea of causation is more intimately tied to generative/productive causation than to counterfactual dependence.

13. See, e.g., "Mechanism, Purpose, and Explanatory Exclusion," reprinted in my *Supervenience and Mind* (Cambridge: Cambridge University Press, 1993); first published in 1989.

supervenience—that is, for anyone who is a minimal physicalist. I have called the line of considerations to be presented below "the supervenience argument"; in the literature, it is also known as "the exclusion argument." (For usage uniformity, it is best to think of the supervenience argument as a special form of the exclusion argument, and take the latter as a generic form of argument with the conclusion that mental cause is always excluded by physical cause.) Briefly, the argument goes like this.[14] Suppose that an instantiation of mental property M causes another mental property, M*, to instantiate. (We take property instantiations as events; instantiations of a mental property are mental events, and similarly for physical properties and physical events.) This is perfectly consistent with physical causal closure. But mind-body supervenience says that this instantiation of mental property M* occurs in virtue of the fact that one of the physical properties on which M* supervenes is instantiated at that time; call this physical base property P*. This means that given that P* is instantiated on this occasion, M* must of necessity be instantiated on this occasion. That is, the M*-instance is wholly dependent on, and is generated by, the P*-instance. At this point, the exclusion principle kicks in: Is the occurrence of the M*-instance due to its supposed cause, the M-instance, or its supervenience base event, P*-instance? It must be one or the other, but which one? Given that its physical supervenience base P* is instantiated on this occasion, M* must be instantiated as well on this occasion, regardless of what might have preceded this M*-instance. In what sense, then, can the M-instance be said to be a "cause," or a generative source, of the M*-instance?

14. This argument will be discussed in greater detail in chapter 2, including responses to some of the objections and criticisms that have been raised against it. I first presented this argument in an explicit form in "'Downward Causation' in Emergentism and Nonreductive Materialism," in *Emergence or Reduction?*, ed. Ansgar Beckermann, Hans Flohr, and Jaegwon Kim (Berlin: De Gruyter, 1992).

I believe that the only acceptable way of reconciling the two causal/generative claims and achieving a consistent picture of the situation is this: the M-instance caused the M*-instance *by causing* the P*-instance. More generally, the following principle seems highly plausible: *In order to cause a supervenient property to be instantiated, you must cause one of its base properties to be instantiated.* In order to alter the aesthetic properties of a work of art, you must alter the physical properties on which the aesthetic properties supervene; in order to do something about your headache you must causally intervene in the brain state on which the headache, supervenes. There is no other way; this is what makes the idea of telepathy (for example, a thought of mine directly causing a thought in you) not credible if not incoherent—unless of course one could telepathically influence another person's brain processes. (In fact, for present purposes, this principle concerning the causation of supervenient properties, which I believe is independently plausible, can replace the principle of determinative/generative exclusion, which some might find too broad.)

So M causes M* to instantiate by causing P* to instantiate, from which it trivially follows that the M-instance causes a P*-instance. But this is a case of mental-to-physical causation. Turning our attention now to the supposed mental cause M, we see that, by mind-body supervenience, M must have its own physical supervenience base; call it P. When we consider the total picture, there seems every reason to consider P to be a cause of P*. If we think of causation in terms of sufficiency, P is clearly sufficient for P*, since it is sufficient for M and M is sufficient for P*. If we think of causation in terms of counterfactuals, we may assume that if P had not been there, the supervening M wouldn't have been there either, and that since M is what brought about P*, P* wouldn't be there either. So at this point we have the following two causal claims: M causes P*, and P causes P*.

Now, given psychophysical property dualism espoused by the nonreductive physicalist, M and P are distinct properties.

This means that P* has two causes each sufficient for it and occurring at the same time (a supervenient property and its base properties are always instantiated at the same time). At this point the causal exclusion principle applies: either M or P must be disqualified as P*'s cause. A moment's reflection shows that it is M that must be disqualified. The reason is that if P is disqualified, the causal closure principle kicks in again, saying that since a physical event, P*, has a cause (namely M), it must have a physical cause (occurring at the same time as M)—the disqualified P will do—and we are back in the same situation, a situation in which we again have to choose between a physical and a mental cause. Unless mental cause M is jettisoned in favor of P, we would be off to an infinite regress—or be forever treading water in the same place.

The final picture that has emerged is this: P is a cause of P*, with M and M* supervening respectively on P and P*. There is a single underlying causal process in this picture, and this process connects two physical properties, P and P*. The correlations between M and M* and between M and P* are by no means accidental or coincidental; they are lawful and counterfactual-sustaining regularities arising out of M's and M*'s supervenience on the causally linked P and P*. These observed correlations give us an impression of causation; however, that is only an appearance, and there is no more causation here than between two successive shadows cast by a moving car, or two successive symptoms of a developing pathology. This is a simple and elegant picture, metaphysically speaking, but it will prompt howls of protest from those who think that it has given away something very special and precious, namely the causal efficacy of our minds. Thus is born the problem of mental causation.

The problem of mental causation. Causal efficacy of mental properties is inconsistent with the joint acceptance of the following four claims: (i) physical causal closure, (ii) causal exclusion,

(iii) mind-body supervenience, and (iv) mental/physical property dualism—the view that mental properties are irreducible to physical properties.

Physical causal closure and mind-body supervenience are, or should be, among the shared commitments of all physicalists. The exclusion principles are general metaphysical constraints, and I don't see how they can be successfully challenged. This leaves mind-body property dualism as the only negotiable item. But to negotiate it away is to embrace reductionism. This will cause a chill in those physicalists who want to eat the cake and have it too—that is, those who want both the irreducibility and causal efficacy of the mental. I believe that the question no longer is whether or not those of us who want to protect mental causation find mind-body reductionism palatable. What has become increasingly clear after three decades of debate is that if we want robust mental causation, we had better be prepared to take reductionism seriously, whether we like it or not. But even if you are ready for reductionism, it doesn't necessarily mean that you can have it. For reductionism may not be true. This is the point to which we now turn.

Can We Reduce Qualia?

Before reduction and reductionism can be usefully discussed, we need to be tolerably clear about the model of reduction appropriate to the issues on hand. I believe much of the philosophical debate during the past few decades concerning the reducibility of the mental has turned out to be a futile exercise because it was predicated on the wrong model of reduction. This is the derivational model of intertheoretic reduction developed by Ernest Nagel in the 1950s and '60s. As is widely known, the heart of Nagel reduction is *bridge laws*, the empirical lawlike principles that are supposed to connect the properties of the domain to be reduced with the properties of the base domain.

Specifically, the requirement, as standardly understood, is that each property up for reduction be connected by a bridge law with a nomologically coextensive property in the base domain. Most of the influential antireductionist arguments—notably, Davidson's anomalist argument and the Putnam-Fodor multiple realization argument[15]—have focused on showing that the bridge law requirement cannot be met for mental properties in relation to physical/biological properties.

All this is by now a familiar story, and there is no need here to rehearse the arguments, counterarguments, and so forth. But the philosophical emptiness of Nagel reduction is quickly seen when we notice that a Nagel reduction of the mental to the physical is consistent with, and even in some cases entailed by, many all-out dualisms, such as the double-aspect theory, the doctrine of preestablished harmony, epiphenomenalism, and even emergentism. The reason of course is that these dualisms are consistent with the mind-body bridge law requirement; in fact, some of them, like the double-aspect theory, entail the satisfaction of this requirement. This objection can be circumvented by strengthening the bridge laws into identities—that is, by requiring the bridging principles connecting the reducing and reduced theories to take the form of an identity ("pain = C-fiber activation") rather than a biconditional law ("pain occurs to an organism at a time just in case its C-fibers are activated at that time")—that is, by moving from bridge-law reduction to identity reduction.[16] It has recently

15. Donald Davidson, "Mental Events," reprinted in his *Essays on Actions and Events* (Oxford and New York: Oxford University Press, 1980); first published in 1970. Hilary Putnam, "The Nature of Mental States," in his *Philosophical Papers*, vol. 2 (Cambridge: Cambridge University Press, 1975); first published in 1967. Jerry A. Fodor, "Special Sciences—or the Disunity of Science as a Working Hypothesis," *Synthese* 27 (1974): 97–115.

16. As early as the 1970s Robert L. Causey argued that microreduction requires cross-level identities of properties, and that genuine reductions cannot be based merely on bridge laws affirming property correlations; see his "Attribute-Identities in Microreductions," *Journal of Philosophy* 69 (1972): 407–422.

been suggested that an identity reduction of consciousness is just what is needed to close the much-discussed "explanatory gap" between the brain and conscious experience. We will look at the feasibility of identity reduction for consciousness in later chapters (chapters 4 and 5). The main problem with this proposal, as we will see, concerns the availability of mind-body identities for reductive purposes. I will argue that the principal arguments advanced for psychoneural identities, namely that they serve certain essential explanatory purposes, do not work, and that there is no visible reason to think that such identities are true or that we will ever be entitled to them.

What then is required to reduce a mental property, say pain? I believe that what has to be done is, first, to *functionalize* pain (or, more precisely, the property of being in pain): namely, to show that being in pain is definable as being in a state (or instantiating a property) that is caused by certain inputs (i.e., tissue damage, trauma) and that in turn causes certain behavioral and other outputs (i.e., characteristic pain behaviors, a sense of distress, a desire to be rid of it). More generally, instantiating a mental property M, upon M's functionalization, will turn out to be being in some state or other that is typically caused by a certain specified set of stimulus conditions and that in turn typically causes a certain specified set of outputs. Next, once a mental property has been functionalized, we can look for its "realizers"—that is, states or properties that satisfy the causal specification defining that mental property. Thus, for pain, we look for an internal state in an organism that is caused to instantiate by tissue damage and trauma and whose instantiation in turn causes characteristic pain behaviors (and possibly outputs of other kinds). In the case of humans and perhaps mammals in general, the state turns out to be, let us say, electrical activity in a certain cortical zone—call it Q. That is, neural state Q is the realizer of pain for humans and mammals. Conventional wisdom has it that pain and other mental states have multiple diverse realizers

across different species and structures, and perhaps even among members of the same species (or even in the same individual over time). This means that this second step of finding realizers of a mental property is likely to be an ongoing affair with no clear end. Obviously, we are not going to find, nor would we necessarily be interested in identifying, all actual and possible realizers of pain for all actual and possible pain-capable organisms and systems. Functional reduction, as I call it, can focus on the reduction of a mental property, or a group of them, for a specific population—that is, neural research on pain will aim at *local* reductions, not a one-shot *global* reduction (as suggested by the Nagel bridge-law model). We may be interested in finding the neural basis of human pain, or canine pain, or Martian pain. We may be interested in identifying the neural basis of your pain now or my pain yesterday. Neural bases may differ for different instances of pain, but individual pains must nonetheless reduce to their respective neural/physical realizers. Unlike in the case of Nagelian bridge-law reduction, the multiple realizability of pain is no barrier to local reduction by functionalization. Suppose that pain has physical realizers, P_1, P_2, ... Then, any given instance of pain is an instance of either P_1 or of P_2 or . . . . If you are in pain in virtue of being in state P_k, there is nothing more, or less, to your being in pain than your being in state P_k. This particular pain is the very same state as this instance of P_k. Each pain instance is a P_1-instance, or P_2-instance, or . . . ; that is, all pain instances reduce to the instances of its realizers.[17]

If pain can be functionalized in this sense, its instances will have the causal powers of pain's realizers. Thus, if a given

17. See my "Making Sense of Emergence," *Philosophical Studies* 95 (1999): 3–36, and *Mind in a Physical World* (Cambridge, MA: MIT Press, 1998) for more details, in particular concerning how reductions conforming to this model meet the basic methodological and metaphysical requirements of reduction. More details on functional reduction can be found in chapter 4 below.

instance of pain occurs in virtue of the instantiation of physical realizer P_k, that pain instance has the causal powers of this instance of P_k. This will solve the problem of the causal efficacy of pains—that is, provided that pain can be functionalized. It is important to see that this result cannot be achieved by simply assuming that P_k is a *neural correlate*, or *substrate*, of pain. It might be that pain and P_k correlate with each other because they are both the effects of a common cause; if such is the case there obviously is no reason for thinking that a given occurrence of pain and the corresponding instance of P_k have the same causal powers, or that they are one and the same event. Pain and its realizers are much more intimately related: to be in pain is to be in a state meeting causal specification C—that is, to be in pain *is* to instantiate one of its realizers—and if you are in pain in virtue of instantiating pain-realizer P_k, there is no pain event over and above this instantiation of P_k.

So if pain is functionalized, the problem of mental causation has a simple solution for all pain instances. But what of the causal efficacy of pain itself? What should we say about the causal powers of pain as a mental kind? The answer is that as a kind pain will be causally heterogeneous, as heterogeneous as the heterogeneity of its diverse realizers. Pain, as a kind, will lack the kind of causal/nomological unity we expect of true natural kinds, kinds in terms of which scientific theorizing is conducted. This is what we must expect given that pain is a functional property with multiple diverse physical realizers. If the term "multiple" in "multiple realizations" means anything, it must mean causal/nomological multiplicity; if two realizers of pain are not causally or nomologically diverse, there is no reason to count them as two, not one. On this reductive account, pain will not be causally impotent or epiphenomenal; it is only that pain is causally heterogeneous.

The key question then is this: Is pain functionally reducible? Are mental properties in general functionalizable and hence

functionally reducible? Or are they "emergent" and irreducible? I believe that there is reason to think that intentional/cognitive properties are functionalizable. However, I am with those who believe that phenomenal properties of consciousness are not functional properties. To argue for this view of phenomenal properties, or qualia, we do not need anything as esoteric and controversial as the "zombie" hypothesis much discussed recently[18]—that is, the claim that zombies, creatures that are indiscernible from us physically and behaviorally but who lack consciousness, are metaphysically possible. All we need is something considerably more modest, namely the metaphysical possibility of qualia inversion. Perhaps the problem is still open, but I believe there are substantial and weighty reasons, and a sufficiently broad consensus among the philosophers who work in this area,[19] to believe that qualia are functionally irreducible.

Moreover, it is easily seen that if qualia are functionally reducible, the problem posed by James and others about consciousness can be solved. Suppose that pain has been functionalized and its realizer identified for humans. Consider a functional characterization of pain like this: To be in pain is to be in a state that is caused by tissue damage and that in turn causes winces and groans. And assume that the venerable C-fiber stimulation is the neural realizer of pain in humans. Consider now the question: Why is Jones in pain at *t*? Can we derive the statement "Jones is in pain at *t*" from information exclusively about Jones's physical/behavioral properties (along with other strictly physical/behavioral information)? Given the functional

18. See David Chalmers, *The Conscious Mind* (Oxford and New York: Oxford University Press, 1996).

19. To mention a few: Ned Block, Christopher Hill, Frank Jackson, Joseph Levine, Colin McGinn, and Brian McLaughlin. Issues mentioned in this paragraph will be discussed in greater detail in the chapters to follow.

reduction, the answer is yes, as is shown by the following deduction:

Jones's C-fibers are stimulated at *t*.

C-fiber stimulation (in humans) is caused by tissue damage and it in turn causes winces and groans.

To be in pain, by definition, is to be in a state which is caused by tissue damage and which in turn causes winces and groans.

Therefore, Jones is in pain at *t*.

Notice that the third line, a functional definition of pain, does not represent empirical/factual information about pain; if anything, it gives us information about the concept pain, or the meaning of "pain." Formally, definitions do not count as premises of a proof; they come free. Notice, moreover, that the displayed derivation could also serve as a prediction of Jones's pain from physical/behavioral information alone. And we could easily convert it into an explanation of why (in humans) pain correlates with C-fiber stimulation, not with another neural state.[20] This derivation would, therefore, answer William James's question why sensations "accompany the brain's workings," a question for which he saw "no glimmer of an explanation." Functional reduction of pain and other sensations would deliver the explanation James was seeking. The only problem is that sensations, or qualia, resist functional reduction, and, as James says, there still is no glimmer of an explanation. But we have made some progress: we now know what is needed to achieve such an explanation.

As earlier noted, there are those who think that functional reduction is not the only way to solve the problem of consciousness; they argue that although pain and other qualia may not be functionally reducible, they are reducible in another way,

20. These issues will be discussed in more detail in chapter 4.

through their identification with physical/neural properties, and that this will enable us to close the gap between consciousness and the brain and thereby provide us with an answer to James's question. We will see in later chapters why this new mind-brain identity reduction is not an option for us. As we will argue,[21] if functional reduction doesn't work for qualia, nothing will.

The Two World-Knots

Let us take stock of where we are: the problem of mental causation is solvable for a given class of mental properties if and only if these properties are functionally reducible with physical/biological properties as their realizers. But phenomenal mental properties are not functionally definable and hence functionally irreducible. Hence, the problem of mental causation is not solvable for phenomenal mental properties.

But, as we also saw, the problem of consciousness, or "the mystery of consciousness," is solvable if consciousness is functionally reducible—and I will argue that it is solvable *only* if consciousness is functionally reducible. So the functional irreducibility of consciousness entails the unsolvability of both the problem of consciousness and the problem of mental causation—at least as the latter problem concerns consciousness. It is thus that the two problems, that of mental causation and that of consciousness, turn out to share an interlocking fate. What stands in the way of solving the problem of mental causation is consciousness. And what stands in the way of solving the problem of consciousness is the impossibility of interpreting or defining it in terms of its causal relations to physical/biological properties. They are indeed *Weltknoten*, problems that have eluded our best philosophical efforts. They seem deeply

21. In chapters 4 and 5.

entrenched in the way we conceptualize the world and ourselves, and seem to arise from some of the fundamental assumptions we hold about each.

Does this mean that there is some hidden flaw somewhere in our system of concepts and assumptions, and that we need to alter, in some basic way, our conceptual framework to rid ourselves of these problems? Of course, if our scheme of concepts were radically altered, the problems would be altered as well; perhaps, the new scheme would not even permit these, or equivalent, problems to be formulated. Some philosophers would be willing to take this as a sufficient ground for urging us to abandon our present system of concepts in favor of a cleansed and tidier one, claiming that the conundrum of mental causation and consciousness is reason enough for jettisoning our shared scheme of intentional and phenomenal idioms, with its alleged built-in "Cartesian" errors and confusions. There are others who blame our penchant for thinking in terms of robust productive causality for the vexing problem of mental causation. Blaming our system of concepts, or our language, for philosophical difficulties is a familiar philosophical strategy of long standing. To me, this often turns out to be an ostrich strategy—trying to avoid problems by ignoring them. To motivate the discarding of a framework, we need independent reasons—we should be able to show it to be deficient, incomplete, or flawed in some fundamental way, independently of the fact that it generates puzzles and problems that we are unable to deal with. Why should we suppose that all problems are solvable—and solvable by us? (Just because we find difficult, perhaps insoluble, moral problems and puzzles, should we cast aside moral concepts and moral discourse?) It may well be that our mind-body problem, or something close to it, arises within any scheme that is rich enough to do justice to the world as we experience it. It may well be that the problem is an inexorable consequence of the tension between the objective world of physical existence and the

subjective world of experience, and that the distinction between the objective and the subjective is unavoidable for reflective cognizers and agents of the kind that we are.[22]

To conclude, then, the mind-body problem, for us, the would-be physicalists, has come down to two problems, mental causation and consciousness, and these together represent the most profound challenge to physicalism. If physicalism is to survive as a worldview for us, it must show just where we belong in the physical world, and this means that it must give an account of our status as conscious creatures with powers to affect our surroundings in virtue of our consciousness and mentality. The arguments that have been presented here already suggest that physicalism will not be able to survive intact and in its entirety. We will try to determine how much of it can survive, and we will see, I hope, that what does survive is good enough for us.

22. A thought like this is suggested by Thomas Nagel in *The View from Nowhere* (New York: Oxford University Press, 1986).

2

The Supervenience Argument Motivated, Clarified, and Defended

AN ARGUMENT was presented in the preceding chapter to show that, on an influential position on the mind-body problem, mental properties turn out to be without causal efficacy. This is what I have called the supervenience argument, also called the exclusion argument in the literature. The argument has drawn comments, criticisms, and objections from a wide range of philosophers, but mostly from those who want to defend orthodox nonreductive physicalism and other forms of mind-body property dualism. Critics of the argument have raised some significant issues, both about the specifics of the argument and, more interestingly, about the broader philosophical issues involved. In this chapter, I would like to address two of the more pressing problems. One is that of "overdetermination," brought up by a number of philosophers; the second is the problem of "causal drainage," forcefully developed by Ned Block in his "Do Causal Powers Drain Away?"[1] Before we get to these and other issues, I want to set out the leading idea that motivates the supervenience argument and then offer what

1. Ned Block, "Do Causal Powers Drain Away?" *Philosophy and Phenomenological Research* 67 (2003): 133–150.

I hope will be a clearer statement of the argument, along with explanatory comments that some may find useful. But first we need a brief description of the philosophical position that is the target of the supervenience argument.

Nonreductive Physicalism

There is no consensus on exactly how nonreductive physicalism is to be formulated, for the simple reason that there is no consensus about either how physicalism is to be formulated or how we should understand reduction. For present purposes, however, no precise formulation is needed; a broad-brush characterization will be sufficient. Moreover, there need not be a single "correct" or "right" formulation of physicalism; there probably are a number of claims, not strictly equivalent, about the fundamentally physical character of the world, each of which can reasonably be considered a statement of physicalism. The strengths and weaknesses, merits and demerits, of these different physicalisms could be examined and debated, and reasonable people could come to different conclusions about them. In any case, most will agree that the following three doctrines are central to nonreductive physicalism: mind-body supervenience, the physical irreducibility of the mental, and the causal efficaciousness of the mental. Mind-body supervenience, the claim that makes the position a form of physicalism, can be stated as follows:

> *Supervenience.* Mental properties strongly supervene on physical/biological properties. That is, if any system *s* instantiates a mental property M at *t*, there necessarily exists a physical property P such that *s* instantiates P at *t*, and necessarily anything instantiating P at any time instantiates M at that time.[2]

2. There are alternative, not quite equivalent, ways of stating mind-body supervenience; one could get a good idea of what these might be from Brian McLaughlin, "Varieties of Supervenience," in *Supervenience: New Essays*, ed. Elias

I take supervenience as an ontological thesis involving the idea of dependence—a sense of dependence that justifies saying that a mental property is instantiated in a given organism at a time *because*, or *in virtue of* the fact that, one of its physical "base" properties is instantiated by the organism at that time. *Supervenience*, therefore, is not a mere claim of covariation between mental and physical properties; it includes a claim of existential dependence of the mental on the physical. I am assuming that a serious physicalist will accept this interpretation of supervenience. Mind-body supervenience as a bare claim about how mental and physical properties covary will be accepted by the double-aspect theorist, the neutral monist, the emergentist, and the epiphenomenalist; it can be accepted even by the substance dualist.

The second component of nonreductive physicalism reflects the "nonreductive" character of this form of physicalism:

> *Irreducibility.* Mental properties are not reducible to, and are not identical with, physical properties.

There is no single well-defined sense, or model, of reduction shared by all disputants in this debate, but this will not matter for us in the context of the supervenience argument; all we need to assume here is that physically irreducible properties remain outside the physical domain—that is, if anything is physically reduced, it must be identical with some physical item. The root meaning of reduction was given, I believe, by J.J.C. Smart when he said that sensations are nothing "over and above" brain processes.[3] If Xs are reduced to Ys, then Xs are nothing over and above the Ys.

Savellos and Ümit Yalçin (Cambridge: Cambridge University Press, 1995). In some contexts the interpretation of "necessarily" as it occurs in the last clause can be crucial; for our purposes, there is no need to opt for any special specification.

3. J.J.C. Smart, "Sensations and Brain Processes," in *The Nature of Mind*, ed. David M. Rosenthal (New York and Oxford: Oxford University Press, 1991), p. 170. Originally published in *Philosophical Review* 68 (1959): 141–56.

We now come to the third doctrine, concerning the causal status of these irreducible mental properties.

> *Causal efficacy.* Mental properties have causal efficacy—that is, their instantiations can, and do, cause other properties, both mental and physical, to be instantiated.

This last thesis is important to the many friends of the position I am describing. The irreducibility claim is often motivated by a desire to save mental properties as something special and distinctive, but if these properties turn out to be causally impotent and explanatorily useless, that would rob them of any real interest or significance, rendering the issue of their reducibility largely moot. Or one could argue that since physical properties are assumed to be causally efficacious, causally inert mental properties obviously cannot be physically reduced. This means that the rejection of mental causal efficacy would make the irreducibility claim true but trivial. In these ways, therefore, the doctrines of irreducibility and causal efficacy go hand in hand.

It can be debated whether these three doctrines constitute a robust enough physicalism. The issue obviously turns on the question whether mind-body supervenience as stated is sufficient for physicalism, since the irreducibility and mental causal efficacy have nothing specifically to do with physicalism; Descartes endorsed both. Moreover, classic emergentism, not usually considered a form of physicalism, endorsed all three, making it a target of the supervenience argument.[4] However, this issue will not affect the discussions to follow. My claims and arguments are intended to apply to any position that accepts the three propositions; what else it accepts makes no difference.

4. See my "Being Realistic about Emergence" in *The Emergence of Emergence*, ed. Paul Davies and Philip Clayton (Oxford: Oxford University Press, forthcoming). The three doctrines, however, can be thought of as capturing the physicalist core of emergentism. On supervenience and physicalism, see Jessica Wilson, "Supervenience-Based Formulations of Physicalism," forthcoming in *Noûs*.

The Fundamental Idea

The idea that drives the supervenience argument can be expressed in the following proposition, which I name after the great eighteenth-century American theologian-philosopher Jonathan Edwards:

> *Edwards's dictum.* There is a tension between "vertical" determination and "horizontal" causation. In fact, vertical determination excludes horizontal causation.

What do I mean by "vertical" determination? Consider an object, say this lump of bronze. At any given time it has a variety of intrinsic properties, like color, shape, texture, density, hardness, electrical conductivity, and so on. Most of us would accept the proposition that the bronze has these properties at this time in virtue of the fact that it has, at this time, a certain microstructure—that is, it is composed of molecules of certain kinds (copper and tin) in a certain specific structural configuration. I describe this situation by saying that the macroproperties of the bronze are vertically determined by its synchronous microstructure. The term "vertical" is meant to reflect the usual practice of picturing micro-macro levels in a vertical array, with the micro underpinning the macro. In contrast, we usually represent diachronic causal relations on a horizontal line, from past (left) to future (right)—"time's arrow" seems always to fly from left to right. From the causal point of view, the piece of bronze has the properties it has at t because it had the properties it had at $t - \Delta t$ (and certain boundary conditions obtained during this period). The past determines the future and the future depends on the past. That is what I mean by "horizontal" causation. So we have here two purported determinative relationships orthogonal to each other: vertical micro-macro mereological determination and horizontal past-to-future causal determination.

The lump of bronze has the color yellow at time t. Why is it yellow at t? There are two presumptive answers: (1) because its

surface has microstructural property M at t; (2) because it was yellow at $t - \Delta t$. To appreciate the force of the supervenience argument it is essential to see a prima facie tension between these two explanations. As long as the lump has microproperty M at t, it's going to be yellow at t, *no matter what happened before t*. Moreover, unless the lump has M, or another appropriate microproperty (with the right reflectance characteristic), at t, it cannot be yellow at t. Anything that happened before t seems irrelevant to the lump's being yellow at t; its having M at t is fully sufficient in itself to make it yellow at t.

So far as I know, Jonathan Edwards was the first philosopher who saw a tension of precisely this kind. Edwards' surprising doctrine that there are no temporally persisting objects was based on his belief that the existence of such objects is excluded by the fact that God is the sustaining cause of the created world at every instant of time. There are no persisting things because at every moment God creates, or recreates, the entire world *ex nihilo*—that is what it means to say that God is the sustaining cause of the world. Consider two successive "time slices" of the bronze: each slice is created by God, and there is no causal or other direct existential relationship between them. To illustrate his argument, Edwards offers a marvelously apt analogy:

> The *images* of things in a glass, as we keep our eye upon them, seem to remain precisely the same, with a continuing, perfect identity. But it is known to be otherwise. Philosophers well know that these images are constantly renewed, by the impression and reflection of *new* rays of light; so that the image impressed by the former rays is constantly vanishing, and a *new* image is impressed by *new* rays every moment, both on the glass and on the eye. . . . And the new images being put on *immediately* or *instantly* do not make them the same, any more than if it were done with the intermission of an *hour* or a *day*. The image that exists at this moment is not at all *derived* from the image which existed at the last preceding moment. As may

> be seen, because if the succession of new *rays* be intercepted, by something interposed between the object and the glass, the image immediately ceases; the *past existence* of the image has no influence to uphold it, so much as for a moment.[5]

Successive images are not causally related to each other; they are each caused by something else. If we suppose that the persistence of an object requires causal relations between its earlier and later stages, Edwards is arguing that "horizontal" causation involving created substances is excluded by their "vertical" dependence on God as a sustaining cause of the world at every instant. Remove God as the sustaining cause; the whole world will vanish at that very instant.[6]

It is simple to see how Edwards's dictum applies to the mind-body case, causing trouble for mental causation. Mind-body supervenience, or the idea that the mental is physically "realized"—in fact, any serious doctrine of mind-body dependence will do—plays the role of vertical determination or dependence, and mental causation, or any "higher-level" causation, is the horizontal causation at issue. The tension between vertical determination and horizontal causation, or the former's threat to preempt and void the latter, has been, at least for me, at the heart of the worries about mental causation.

5. Jonathan Edwards, *Doctrines of Original Sin Defended* (1758), Part IV, Chapter II. The quotation is from *Jonathan Edwards*, ed. C. H. Faust and T. H. Johnson (New York: American Book Co., 1935), p. 335. (Italics in the original.) It seems, however, that Edwards's argument may well have been foreshadowed by the occasionalists of the 17th century.

6. Some will argue that these considerations—and some of the crucial steps in the supervenience argument—depend on the use of a robust, "thick" concept of productive or generative causation rather than a "thin" concept based on the idea of counterfactual dependence or simple Humean "constant conjunctions," and that thin causation is all the causation that there is. See Barry Loewer's "Comments on Jaegwon Kim's *Mind in a Physical World*," *Philosophy and Phenomenological Research* 65 (2002): 655–62, and my reply to Loewer, ibid., 674–77.

The Supervenience Argument Refined and Clarified

Let us now turn to a restatement of the supervenience argument in a more explicit and streamlined form. It is useful to divide the argument into two stages; I believe each stage has its own interest, and this will also enable me to present two materially different ways of completing the second stage of the argument.

Stage 1

We begin with the supposition that there are cases of mental-to-mental causation. Let M and M* be mental properties:

(1) M causes M*.

Properties as such don't enter into causal relations; when we say "M causes M*," that is short for "An instance of M causes an instance of M*" or "An instantiation of M causes M* to instantiate on that occasion." Also for brevity we suppress reference to times. From *Supervenience*, we have:

(2) For some physical property P*; M* has P* as its supervenience base.

As earlier noted, (1) and (2) together give rise to a tension when we consider the question "Why is M* instantiated on this occasion? What is responsible for, and explains, the fact that M* occurs on this occasion?" For there are two seemingly exclusionary answers: (a) "Because M caused M* to instantiate on this occasion," and (b) "Because P*, a supervenience base of M*, is instantiated on this occasion." This of course is where Jonathan Edwards's insight, encapsulated in Edwards's dictum, comes into play: Given that P* is present on this occasion, M* would be there no matter what happened before; as M*'s supervenience base, the instantiation of P* at *t* in and of itself

necessitates M*'s occurrence at *t*. This would be true even if M*'s putative cause, M, had not occurred—*unless, that is, the occurrence of M had something to do with the occurrence of P* on this occasion*. This last observation points to a simple and natural way of dissipating the tension created by (a) and (b):

(3) M caused M* *by* causing its supervenience base P*.

This completes Stage 1. What the argument has shown at this point is that if *Supervenience* is assumed, mental-to-mental causation entails mental-to-physical causation—or, more generally, that "same-level" causation entails "downward" causation. Given *Supervenience*, it is not possible to have causation in the mental realm without causation that crosses into the physical realm. This result is of some significance; if we accept, as most do, some doctrine of macro-micro supervenience, we can no longer isolate causal relations within levels; any causal relation at level L (higher than the bottom level) entails a cross-level, L to L − 1, causal relation. In short, *level-bound causal autonomy is inconsistent with supervenience or dependence between the levels*. Further, an important part of the interest of the supervenience argument is that it shows that, under the physicalist assumptions we are working with, mind-to-mind causation is in trouble just as much as mind-to-body causation. Often the problem of mental causation is presented as that of explaining how the mental can inject causal influences into the causally closed physical domain, that is, the problem of explaining mental-to-physical causation. I wanted to do something more, namely to show that physicalism can put in peril all forms of mental causation, including mental-to-mental causation.[7] This is why the argument begins with line (1). It is at Stage 2 that we take up mental-to-physical causation. It is noteworthy that,

7. As we will see in the next chapter, an interesting parallel holds in the case of substance dualism: under substance dualism, mental-to-mental causation turns out to be as problematic as mental-to-physical causation.

unlike in the second stage below, the argument up to this point makes no explicit appeal to any special metaphysical principles; in particular, no specific assumptions about the physical domain, such as its causal closure or completeness, enter the picture at this stage.[8] Mental-physical supervenience is the only substantive premise that has been in play thus far.

Stage 2

There are two ways of completing the argument, and I believe the second, which is new, is of some interest. I will first present the original version in a somewhat clearer form:

COMPLETION I

We now turn our attention to M, the supposed mental cause of M*. From *Supervenience*, it follows:

(4) M has a physical supervenience base, P.

There are strong reasons for thinking that P is a cause of P*. I will not rehearse the considerations in support of this idea; let us just note that P is (at least) nomologically sufficient for M, and the occurrence of M on this occasion depends on, and is determined by, the presence of P on this occasion. Since ex hypothesi M is a cause of P*, P would appear amply to qualify as a cause of P* as well. So we have:

(5) M causes P*, and P causes P*.

8. On some occasions I have tried to argue for (3) by invoking an exclusion principle—see, for example, the "principle of determinative/generative exclusion" in chapter 1. I think it preferable not to appeal to any general principle here; I now prefer to rely on the reader's seeing the tension I spoke of in connection with the two answers to the question "Why is M* instantiated on this occasion?" Anyone who understands Jonathan Edwards's argument and his mirror analogy will see it; I don't believe invoking any "principle" will help persuade anyone who is not with me here.

Note that P's causation of P* cannot be thought of as a causal chain with M as an intermediate causal link; one reason is that the P-to-M relation is not a causal relation. Note also that since M supervenes on P, M and P occur precisely at the same time. (Moreover, as we will shortly see, the two principles that will be introduced, *Exclusion* and *Closure*, together disqualify M as a cause of P*, making the idea of a causal chain from P to M to P* a nonstarter.)

To continue, from *Irreducibility*, we have:

(6) M ≠ P.[9]

Again, (5) and (6) present to us a situation with metaphysical tension. For P* is represented here as having two distinct causes, each sufficient for its occurrence. The situation is ripe for the application of the causal exclusion principle, which can be stated as follows:

> *Exclusion*. No single event can have more than one sufficient cause occurring at any given time—unless it is a genuine case of causal overdetermination.

Let us assume that this is not a case of causal overdetermination (we will discuss the overdetermination issue below).

(7) P* is not causally overdetermined by M and P.

By *Exclusion*, therefore, we must eliminate either M or P as P*'s cause. Which one?

9. Note: this only means that this instance of M ≠ this instance of P. Does this mean that a Davidsonian "token identity" suffices here? The answer is no: the relevant sense in which an instance of M = an instance of P requires either property identity M = P or some form of reductive relationship between them. (See *Mind in a Physical World*, ch. 4). The fact that properties M and P must be implicated in the identity, or nonidentity, of M and P instances can be seen from the fact that "An M-instance causes a P-instance" must be understood with the proviso "in virtue of the former being an instance of M and the latter an instance of P."

(8) The putative mental cause, M, is excluded by the physical cause, P. That is, P, not M, is a cause of P*.

We can give relatively informal reasons for choosing P over M as the cause of P*, but for a general theoretical justification we may appeal to the causal closure of the physical domain:

> *Closure*. If a physical event has a cause that occurs at *t*, it has a physical cause that occurs at *t*.[10]

If we were to choose M over P as P*'s cause, *Closure* would kick in again, leading us to posit a physical cause of P*, call it P_1 (what could P_1 be if not P?), and this would again call for the application of *Exclusion*, forcing us to choose between M and P_1 (that is, P). Unless P is chosen and M excluded, we would be off to an unending repetition of the same choice situation; M must be excluded and P retained.

It is worthwhile to reflect on how *Exclusion* and *Closure* work together to yield the epiphenomenalist conclusion (8). *Exclusion* itself is neutral with respect to the mental-physical competition; it says either the mental cause or the physical cause must go, but doesn't favor either over the other. What makes the difference—what introduces an asymmetry into the situation—is *Closure*. It is the causal closure of the physical world that excludes the mental cause, enabling the physical cause to prevail. If the situation with causal closure were the reverse, so that it was the mental domain, not the physical domain, that was causally closed, the mental

10. For discussion of physical causal closure, or "completeness," see, e.g., David Papineau, *Thinking about Consciousness* (Oxford: Clarendon Press, 2002), ch. 1; E. J. Lowe, "Physical Causal Closure and the Invisibility of Mental Causation," in *Physicalism and Mental Causation*, ed. Sven Walter and Heinz-Dieter Heckmann (Exeter, UK: Imprint Academic, 2003). A simpler statement of causal closure in the form "If a physical event has a cause, it has a physical cause" will not do; given the transitivity of causation, the requirement would be met by a causal chain consisting of a physical effect caused by a mental cause which in turn is caused by a physical cause.

cause would have prevailed over its physical competitor. I suppose this could happen under some forms of Idealism; one would then worry about the "problem" of physical causation.

COMPLETION 2

Let us begin with the last line of Stage 1:

(3) M causes M* by causing its physical supervenience base P*.

From which it follows:

(4) M is a cause of P*.

By *Closure* it follows:

(5) P* has a physical cause—call it P—occurring at the time M occurs.

(6) M ≠ P (by *Irreducibility*).

(7) Hence, P* has two distinct causes, M and P, and this is not a case of causal overdetermination.

(8) Hence, by *Exclusion*, either M or P must go.

(9) By *Closure* and *Exclusion*, M must go; P stays.

This is simpler than Completion 1. *Supervenience* is not needed as a premise, and the claim that M's supervenience base P has a valid claim to be a cause of P* has been bypassed, making it unnecessary to devise an argument for it. However, Completion 1, in some ways, is more intuitive; it better captures Jonathan Edwards's fundamental insight and makes it particularly salient how putative higher-level causal relations give way to causal processes at a lower level. Either way, the main significance of Stage 2 lies in what it shows about the possible hazards involved in the idea of "downward" causation, namely that *the assumptions of causal exclusion and lower-level causal closure disallow downward causation.*

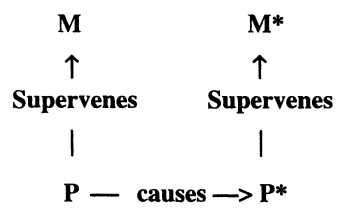

Figure 1.

Figure 1 pictures the outcome of the argument under Completion 1. In this picture, there is but one causal relation, from physical property P to another physical property P*, and the initially posited causal relation from M to M* has been eliminated. An apparent causal relation between the two mental properties is explained away by their respective supervenience on two physical properties that are connected by a genuine causal process. In this picture neither M nor M* is implicated in any causal relations; they play no role in shaping the causal structure—they only supervene on properties that constitute that structure. The supervenience relations together with the causal relation involved can generate counterfactual dependencies between the two mental properties, and between them and the physical properties; but these are no more causal than counterfactual dependencies involving any other supervenient property and its subvenient base (compare the aesthetic properties of a work of art and their base physical properties). Completion 2 presents a picture that is a bit less full: we no longer have the vertical "supervenience" arrow from P to M. M of course must have a physical supervenience base, but the argument, unlike in Completion 1, does not require it to be a cause of P*, although, as Completion 1 suggests, it may well be. The moral, however, is the same: the $M \rightarrow M^*$ and $M \rightarrow P^*$ causal relations have given way to an underlying physical causal process, $P \rightarrow P^*$.

Is Overdetermination an Option?

Several critics have taken issue with line (7), in both completions of the argument, where the claim is made that we should not think of M and P as two distinct overdetermining causes of P*. One thing I said to defend this claim in *Mind in a Physical World* was this: taking the overdetermination option would be in violation of *Closure*, for in a world in which P does not occur but which is as close to the actual world as possible, M would be a cause of P*, leaving P* without a physical cause. My critics have convinced me that what I said there is not quite right and at best incomplete.

Ned Block asks whether in the supposed possible world, one in which the supervenience base P of M does not occur, M could be thought of as occurring at all. If we take away the supervenience base of M, shouldn't that also take away M? This is something to think about. If what Block has in mind is that the following counterfactual may well be true, I agree:

(C) If P had not occurred, M would not have occurred.

For we are apt to reason like this: M was there because P was there, so take away P and M goes as well. "If the patient's nociceptive neurons had not been stimulated at *t*, he would not have experienced pain at *t*," uttered, say, when we deliberately activated these neurons in an experimental situation, would evidently be true. In considering the claim that M and P are each a sufficient cause of P*, however, we need to be able to consider a possible situation in which M occurs without P and evaluate the claim that in this possible situation P* nonetheless follows. If such is not a possible situation—that is, if of necessity any nonP-world is ipso facto a nonM-world—what significance can we attach to the claim that P and M are each an overdetermining sufficient cause of P*, that in addition to P, M also is a sufficient cause of P*? *Supervenience* does not render a nonP-world in

which M occurs impossible; all that *Supervenience* requires is that such a world must include an alternative physical base of M.

So suppose W is a world in which M occurs but P does not. In an instructive and helpful discussion, Thomas Crisp and Ted Warfield have the following to say about such worlds:

> Consider though: either [*Supervenience*] holds in W or it does not. Suppose it does. It follows that M has a physical supervenience base P′ in W. What is the causal status of P′ vis-à-vis P* in W? We won't repeat ourselves, but we saw above an argument of Kim's to the effect that if P′ is a supervenience base for M and M causes P*, then P′ is also causally sufficient for P*. If [*Supervenience*] holds in W, therefore, P* does have a physical cause in W, and [*Closure*] therefore does not fail in W.[11]

Crisp and Warfield are right. Notice, though, that in W, we have a replay of exactly the same situation with which we began—M has a physical base, P′, threatening to preempt it as a cause of P*. In any world in which *Supervenience* holds and M causes P*, some physical property, instantiated at the same time, can claim to be a sufficient cause of P*. As long as *Supervenience* is held constant, there is no world in which M by itself, independently of a physical base, brings about P*; whenever M claims to be a cause of P*, there is some physical property waiting to claim at least an equal causal status. In the actual world, we may suppose that a continuous causal chain connects P with P* (in some cases we may already have detailed neurophysiological knowledge of the physical causal process leading from P to P*).[12] And it would be incoherent to suppose there is another

11. Thomas M. Crisp and Ted A. Warfield, "Kim's Master Argument," *Noûs* 35 (2001): 304–16 (the quoted passage appears on p. 314).

12. In introducing consideration of causal chains, I am implicitly asking the reader to think of causation in terms of actual productive/generative mechanisms involving energy flow, momentum transfer, and the like, and not merely in terms of counterfactual dependencies. Needless to say, the overdetermination idea makes little sense when causation is understood this way.

causal chain from M to P* that is independent of the causal process connecting P with P*; the only plausible supposition is that if there is a causal path from M to P*, that must coincide with the causal path from P to P*. In W, another causal chain connects P′ with P*, and the M-P* chain must coincide with that, and similarly in other such worlds. To be a cause of P*, M must somehow ride piggyback on physical causal chains—distinct ones depending on which physical property subserves M on a given occasion, in the same world or in other possible worlds. And we may ask: In virtue of what relation it bears to physical property P does M earn its entitlement to a free ride on the causal chain from P to P* and to claim this causal chain to be its own? Obviously, the only significant relation M bears to P is supervenience. But why should supervenience confer this right on M? The fact of the matter is that there is only one causal process here, from P to P*,[13] and M's supposed causal contribution to the production of P* is totally mysterious. In standard cases of overdetermination, like two bullets hitting the victim's heart at the same time, the short circuit and the overturned lantern causing a house fire, and so on, each overdetermining cause plays a distinct and distinctive causal role. The usual notion of overdetermination involves two or more separate and independent causal chains intersecting at a common effect. Because of *Supervenience*, however, that is not the kind of situation we have here. In this sense, this is not a case of genuine causal overdetermination, and *Exclusion* applies in a straightforward way. Moreover, anyone tempted by the idea that mental events make their causal contributions by being

13. Some have suggested that the M-to-P* causation is a higher-level "redescription" of the causal process from P to P*. E.g., John R. Searle, "Consciousness, the Brain and the Connection Principle: A Reply," *Philosophy and Phenomenological Research* 55 (1995): 217–32, especially 218–19. Obviously, the redescription strategy is available only to those who accept "M = P," namely reductionist physicalists (Searle of course does not count himself among them).

overdetermining causes should reflect on whether this option could sufficiently vindicate the causal efficacy of the mental.

Now for the second leg of Crisp and Warfield's dilemma:

> Now suppose that [*Supervenience*] does not hold in W. And suppose further that, just as Kim suggests, M causes P* in W without there being any physical cause of P*. Given these assumptions, [*Closure*] does indeed fail in W. But recall that we have supposed along with Kim that the actual world is a Supervenience-world. It follows from this supposition that W is either nomologically or metaphysically impossible, depending on how we read the relevant modal operator in the formulation of [*Supervenience*]. So if W is a world in which [*Closure*] is violated in the way Kim suggests, W is at least nomologically impossible.
>
> What should nonreductivist fans of overdetermination think about this? Should they give up their view because it implies that [*Closure*] fails in worlds that are nomologically (and maybe even metaphysically) impossible? We can't see why they should.[14]

I think we can set aside the possibility that mind-body supervenience is logically or metaphysically necessary, since such a view is essentially a reductionist view,[15] and we are here considering *Supervenience* as a part of nonreductive physicalism. Let us assume then that *Supervenience* is nomologically necessary, and that it fails in W. So in virtue of violating *Supervenience*, W is nomologically impossible. However, W is nomologically impossible not because some physical law is violated in W but because some mental properties fail to supervene on physical properties—that is, because some psychophysical laws of our world fail in W. So W may well be a physically possible world; in fact, we may stipulate W to be a perfect duplicate of our

14. Crisp and Warfield, "Kim's Master Argument," p. 314.

15. This is not an uncontroversial issue, but we cannot go into it here. And there are independent reasons for thinking that mind-brain supervenience, if it holds, must be construed as nomological, not logical or metaphysical, supervenience.

world in all physical respects, including spacetime structure, basic physical laws, and fundamental particles. Should the physicalist not care whether physical causal closure holds in a world like W? Contrary to what Crisp and Warfield suggest, it seems obvious to me that anyone who cares about physicalism should care very much about *Closure* in W.

A more direct way of ruling out overdetermination as an option is to adopt a stronger form of physical causal closure:

> *Strong closure*. Any cause of a physical event is itself a physical event—that is, no nonphysical event can be a cause of a physical event.[16]

Using this principle as a premise has two significant effects. First, it stops the overdetermination option in its tracks; *Strong closure* by itself disallows mental-to-physical causation. Second, *Strong closure* allows us to dispense with *Exclusion*. We no longer need this principle to exclude M in favor of P as P*'s cause, for the simple reason that *Strong closure*, in conjunction with *Irreducibility*, makes M ineligible as a cause of P*.

How might the supervenience argument go under *Strong closure*? Stage 1 is unaffected. Let's briefly look at how Completion 1 might go with *Strong closure*:

(3) M causes M* by causing P*.

(4) M has a physical supervenience base, P.

(5) M causes P*, and P causes P*.

Up to here, the argument is the same as before; from here the argument can continue as follows:

(6*) For every physical property P, M ≠ P *Irreducibility*.

(7*) M does not cause P* (from (6*) and *Strong closure*).

16. An even stronger form of closure can be obtained by also prohibiting physical events from having mental effects—that is, by disallowing all "mixed" causal chains, chains with both physical and mental events.

(8*) M does not cause M* (from (3)[17] and (7*)).

(9*) P causes P* (from (5)).

The outcome is the same as in the original Completion 1, namely Figure 1. But the argument has been simplified in that *Exclusion* has been dispensed with as a premise.

Is this a reason to prefer *Strong closure* to *Closure*? The answer, I believe, is yes and no. Although the causal exclusion principle has been widely accepted and I believe it is virtually an analytic truth with not much content, some find it problematic, and the fact that *Strong closure* makes *Exclusion* dispensable is a point in its favor. (This need not be taken to mean that the argument is no longer properly called an "exclusion" argument; even though no exclusion principle is used as a premise, the *outcome* of the argument is that mental causal relations are "excluded" by physical causality.) Further, there seems no reason for the physicalist to object to *Strong closure*; so why not trade the two premises, *Closure* and *Exclusion*, for a single premise, *Strong closure*, and in the process defuse the overdetermination issue? I believe, though, that there is a philosophical gain in staying with the weaker closure premise. Adopting *Strong closure* as a premise is like starting your argument with mind-body causation already ruled out, at least for nonreductivists; with *Strong closure* as your starting point, there isn't very much more distance you can go or need to go. Perhaps philosophical arguments never make converts out of those who are already committed to the opposite side; but I believe that it can serve philosophical interest to begin with a set of premises that are individually as weak as possible but which somehow conspire together to yield the desired conclusion. It is better, that is to say, to distribute the burden of defending a conclusion among a set of relatively weak premises than to place it on fewer but individually stronger premises.

17. It is implicit in (3) that this is the *only* way M can cause M*.

The latter strategy is apt to provoke the complaint that the argument begs the question and that it serves no useful purpose. I think we learn something about the issues and desiderata involved and their interplay when we run the supervenience argument with *Closure* rather than *Strong closure*.

The Generalization Argument

My main aim in this chapter is to respond to the argument Block has put forward in the following passage:

> The Exclusion Principle [the thesis that "sufficient causation at one level excludes sufficient causation at another level"] leads to problems about causal powers draining away. Kim discusses a number of such problems, including the following two. First, it is hard to believe that there is no mental causation, no physiological causation, no molecular causation, no atomic causation but only bottom level physical causation. Second, it is hard to believe that there is no causation at all if there is no bottom level of physics.[18]

Why does Block think that if the supervenience argument holds, there will be no physiological causation, no molecular causation, etc. any more than mental causation? Because he subscribes to what is called the "generalization argument"—the idea that the supervenience argument generalizes beyond mind-body causation, with the result that causation at *any* level gives way to causation at the next lower level (if there is one), just as the supposed causation at the mental level gets eliminated in favor of causation at the physical/biological level. Block is not alone here. A number of writers have expressed the view that if the supposed problem of mental causation is a real problem, a parallel problem should arise for all other special

18. Block, "Do Causal Powers Drain Away?" p. 138.

sciences, except causation at the most fundamental physical level.[19] Such a view is often stated against the backdrop of a "layered" model of the domains of science, according to which objects and properties of the world are arrayed in a hierarchy of "levels," with the basic physical particles and their properties at the bottom level and, above it, the levels of atoms, molecules, cells, organisms, and so on, all ordered in an ascending ladder-like structure. It is this hierarchical view of the domains of science that gives meaning to the talk of "higher" and "lower" levels—in regard to sciences, laws, explanations, and the rest.[20]

On a hierarchical picture of levels like this, it is natural to think of mental causation only as a special case of higher-level causation. If the supervenience argument shows causation at the psychological level to be preempted by causation at the biological level, why couldn't the argument be iterated to show biological causation to be preempted by physicochemical causation, and so on down to the fundamental microphysical level? The idea that the argument is generalizable this way gains force from the widely accepted assumption that properties at upper levels are supervenient on lower-level properties, the eponymous premise that plays a crucial role in the argument.

Let me begin my response by pointing out that if indeed the supervenience argument is generalizable, that only shows that

19. This includes Tyler Burge, Robert Van Gulick, and many others. See my *Mind in a Physical World*, ch. 3 for references and discussion. Among other discussions of the generalization argument are Paul Noordhof, "Micro-Based Properties and the Supervenience Argument," *Proceedings of the Aristotelian Society* 99 (1999): 109–114; Carl Gillett, "Does the Argument from Realization Generalize? Responses to Kim," *Southern Journal of Philosophy* 39 (2001): 79–98; Thomas D. Bontly, "The Supervenience Argument Generalizes," *Philosophical Studies* 109 (2002): 75–96.

20. Whether a layered model of this kind can be developed as a comprehensive ontology of the world is a debatable issue. I discuss some of the difficulties with such an approach in "The Layered Model: Metaphysical Considerations," *Philosophical Explorations* 5 (2002): 2–20. See also John Heil, *From an Ontological Point of View* (Oxford: Oxford University Press, 2003), ch. 4.

we have a general philosophical problem on hand, and that it is not necessarily a refutation of the argument. If the argument goes wrong, one would like to know just where and how it goes wrong. Moreover, just saying that there "obviously" are biological causation, physiological causation, and so on isn't very helpful; what has to be shown is that these kinds of "higher-level" causation are irreducible to basic physical causation—namely, that there are these causal relations *in addition to* the underlying physical causal processes. It is important to keep in mind that the supervenience argument assumes among its premises the doctrine of the irreducibility of the mental; this premise is invoked at line (6) in both completions of Stage 2. As may be recalled, the argument begins with the supposition that an instance of a mental property M causes another mental property M* to instantiate (line (1)). Block says that this M-to-M* causal relation is "putative—it is a premise in a *reductio* that Kim will reject."[21] But this is not the full story: there is another premise, the premise of irreducibility (line (6): $M \neq P$), against which a reductio can also be performed. This premise, not the supposed M-to-M* causal relation, has always been my primary target. The real aim of the argument, as far as my own philosophical interests are concerned, is not to show that mentality is epiphenomenal, or that mental causal relations are eliminated by physical causal relations; it is rather to show "either reduction or causal impotence." To put it another way, my aim is to force a choice between the situation depicted in figure 1 and what is pictured in figure 2. In this picture, the $M \rightarrow M^*$ causation remains genuine and real; it is the very same causal relation as $P \rightarrow P^*$; the reduction collapses the two levels into one, and there is here one causal relation, not two. The aim of the supervenience argument is to clarify the options available to the physicalist: If you deem yourself a

21. Block, "Do Causal Powers Drain Away?" p. 134.

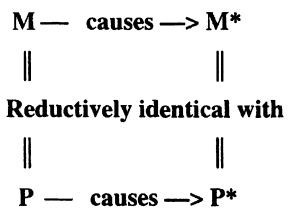

Figure 2.

physicalist, you must choose between figure 1 and figure 2. There are no other options.[22]

Indeed, the supervenience argument may be generalizable, but all that would show is that if there is biological causation, biological properties are, or are reducible to, physical or physicochemical properties; it does not show that biological causation does not exist. The epiphenomenalist brunt of the argument is avoided if one is prepared, and is able, to choose the reductionist branch of the dilemma. It should be kept in mind that merely "choosing" reductionism doesn't make reductionism true; whether or not reductionism is sustainable as an option is an independent question that ought to be decided on its merits.

Many philosophers will reply that biological properties are no more physically reducible than psychological properties, citing their "multiple realizability" in relation to physicochemical properties. For most antireductionist philosophers, multiple realizability has long been a mantra, an all-purpose antireductionist argument applied across the board to all special science properties. They see multiple realization everywhere, and this

22. The underlying metaphysical moral of the two options is the same, however: there is only one causal relation here, namely a physical one, and, more generally, causality is fundamentally a physical phenomenon. An interestingly similar picture results from Donald Davidson's thesis that causation requires "strict laws," and that strict laws are found only in physics. See his "Mental Events," in *Essays on Actions and Events* (Oxford and New York: Oxford University Press, 1980).

leads them to see irreducibility everywhere. I believe, however, that the notion of "realization" as it is often invoked in this context is too loose and ill-formed, and that when realization is properly understood, multiple realization only leads to reducibility to multiple reduction bases, not to irreducibility.[23]

Considerations like those motivating the supervenience argument do not have eliminative implications for macrocausation in general; the supervenience argument does not eliminate all macrocausation, leaving only causal relations between microentities and their properties. This baseball has causal powers that none of its proper parts, in particular none of its constituent microparticles, have, and in virtue of its mass and hardness, the baseball can break a window when it strikes it with a certain velocity. The shattering of the glass was caused by the baseball and certainly not by the individual particles composing it. True, the baseball is a composite object made up of its constituent molecules, atoms, particles, or what have you, and this complex structure consisting of microparticles broke the window. But there is no mystery here: the baseball = this composite structure of microparticles.[24] Presumably, the causal powers of the baseball are *determined* by its microstructural features and perhaps also *explainable* in terms of them. But determination or explanation need have no eliminative implications. Perhaps, macrocausal relations are constituted by, or composed of, a bunch of microcausal relations. But that does not banish macrocausation out of existence any more than the fact that the baseball is composed of microparticles entails its nonexistence. All this is consistent with the supervenience argument.

23. See my "Multiple Realization and the Metaphysics of Reduction," *Philosophy and Phenomenological Research* 52 (1992): 1–26, and *Mind in a Physical World*, ch. 4. For further discussion of multiple realizability and reduction, see John Bickle, *Psychoneural Reduction: The New Wave* (Cambridge, MA: MIT Press, 1998).

24. For a dissenting view—*plus* the view that macrocausation is in general preempted by microcausation—see Trenton Merricks, *Objects and Persons* (Oxford: Clarendon, 2001).

Block's Causal Drainage Argument

A micro-based property of an object is a property characterizing its microstructure—it tells us what sorts of micro-constituents the object is made up of and the structural relations that configure these constituents into a stable object with substantival unity. Micro-based (or microstructural) properties of an object are its macroproperties—they belong to the whole object, not to its constituents—and, moreover, they do not supervene on the properties of the object's micro-constituents. For that reason, the supervenience argument does not touch micro-based properties,[25] and I have claimed that this prevents causal powers from seeping downward from level to level, from macro to micro. Further, I have argued that many chemical and biological properties seem construable as micro-based properties, properties defined or analyzable in terms of microstructure. Block recognizes this as my strategy. The initial criticism he advances can be called the "multiple composition" argument. He writes:

> But why can't micro-based properties be micro-based in *alternative ways*? Why isn't jade an example of a micro-based property, micro-based in both calcium magnesium silicate (nephrite) and sodium aluminum silicate (jadeite)?...
>
> My doubts about [Kim's] picture center on the worry just mentioned concerning multiple decomposition. Micro-based properties are supposed to prevent draining away for both supervenient and functional properties, but Kim's plugging the draining with micro-based properties depends on assuming identities (such as "water = H_2O") and multiple composition will preclude such identities.[26]

25. This has been disputed by some of the authors cited in footnote 19.
26. Block, "Do Causal Powers Drain Away?" pp. 145–46.

Here Block appears to be thinking of multiple composition in parallel with multiple realization: just as multiple realization has been used as an argument against reducibility, multiple composition could be used against identifying a macroproperty, say being jade, with micro-based properties. This is an interesting possibility; multiple compositionality may work as well as multiple realizability, each against its reductionist target. However, I think that neither works very well.

There are two things to say about Block's argument. First, in spite of jade's multiple composition, each instance of jade—that is, each individual piece of jade—is either jadeite or nephrite, and I don't see anything wrong about identifying *its* being jade with *its* being nephrite (if it is nephrite) or with *its* being jadeite (if it's jadeite). If it is nephrite, the causal powers that it has in virtue of being jade will be exactly identical with the causal powers of nephrite. All we need is identity at the level of instances, not necessarily at the level of kinds and properties; causation after all is a relation between property or kind-instances, not between properties or kinds as such. Second, suppose a macroproperty has two or more distinct micro-compositions. We can use the jade example again: we presumably distinguish between the two compositions, jadeite and nephrite, importantly because they are *causally* distinguishable—that is, jadeite and nephrite have significantly different causal profiles. Given this, there are two options. We can either deny that jade is a genuine kind (at least, jade is not a kind of mineral), on account of its causal heterogeneity, or identify jade with a disjunctive kind, jadeite or nephrite (that is, being jade is identified with having the microstructure of jadeite or the microstructure of nephrite). The second option which allows disjunctive kinds is a more conservative approach and may be more viable as a general solution. On the disjunctive approach, being jade turns out to be a causally heterogeneous property, not a causally inert one, and jade turns out to be a causally heterogeneous kind, not a causally irrelevant one. To disarm

Block's multiple composition argument, adopting either disjunctive property/kind identities or instance (or token) identities seems sufficient.

This, however, does not fully block the drainage argument. There may be no causal seepage from macro to micro, but that is not the only way the seepage can occur. The trouble can be seen when we recognize that a given object can have micro-based properties at various levels (the biological, the physicochemical, the atomic, etc.), and that higher-level micro-based properties arguably supervene on their lower-level counterparts. Block has this in mind, I think, when he speaks of "endless subvenience."[27] Other commentators, in particular Ausonio Marras,[28] have also made this point. Let us see how the idea might be developed.

Take any macro-object, O, and let a *total* micro-based property *at level L* be the property corresponding to a complete description of O's microstructure at level L. (Roughly, we can think of "levels" in terms of modes of decomposition of material objects into physically significant constituents; examples of levels are the molecular level, the atomic level, and the level of basic particles.) So if L is the level of the Standard Model, a total micro-based property of O at this level would give a complete description of O's microstructure in terms of the particles and forces posited in the Standard Model. The following is a plausible physicalist principle:

> *Macro-micro supervenience.* All intrinsic properties of O, at any level higher than L, supervene on the total micro-based property of O at level L.

The idea is that wholes made up of the same (qualitatively identical) constituents configured in the same structural relationships

27. Block, "Do Causal Powers Drain Away?" p. 140.

28. Ausonio Marras in "Critical Notice of *Mind in a Physical World*," *Canadian Journal of Philosophy* 30 (2000): 137–60; see p. 151.

will exhibit an identical set of intrinsic properties. Since micro-based properties are intrinsic properties, it follows:

> For any object O, O's micro-based properties at level L supervene on O's total micro-based property at level L*, where L* < L.

Consider a series of total micro-based properties of a given object: M_L, M_{L-1}, M_{L-2}, ... Suppose this series has no end; it continues on, without ever reaching a bottom level. That is, let us suppose that the speculation of the physicists cited by Block is correct, and that matter is infinitely divisible (I will go along with Block that all this makes perfectly good sense; but can we really make sense of the idea of an object that is literally made up of infinitely many physically significant parts, here and now?) According to the supervenience argument, M_L apparently cedes its causal powers to M_{L-1}, whose causal powers in turn are taken over by those of M_{L-2}, and so on without end.

Here, Block's worry appears well placed. The supervenience argument implies the following general proposition:

> *Seepage.* If property Q supervenes on a property Q* at a lower level without being reducible to it, Q's causal powers are preempted by those of Q*.

This means that no member of the infinite series of total micro-based properties M_L, M_{L-1}, . . . has causal powers, since every member has a lower member on which it supervenes. If no member of this series has causal powers, there are none to be had anywhere in the series. Moreover, since all intrinsic properties of the object in question are assumed to supervene on its total micro-based properties at lower levels, none of the object's intrinsic properties can have causal powers, and that means that the object itself has no causal powers. All this on the premise that microphysics has no bottom level and matter is infinitely divisible.[29]

29. For an interesting (skeptical) discussion of the existence of a bottom level, see Jonathan Schaffer, "Is There a Fundamental Level?" *Noûs* 37 (2003): 498–517.

This, I believe, is Block's argument, or at least it is a close-enough approximation to it. As Marras has pointed out, it seems possible to develop the generalization argument within a single level in the micro-macro hierarchy. In any case, the argument is worth thinking about. Compare *Seepage* with the following alternative ways of conceiving the interlevel causal relationship:

> *Explanation.* If property Q supervenes on a property Q* at a lower level without being reducible to it, Q's causal powers (and the causal relations into which Q enters) can be *explained* in terms of the causal powers of Q*.
>
> *Constitution.* If Q supervenes on Q*, Q's causal powers are *constituted* by those of Q*.
>
> *Derivation/determination.* If Q supervenes on Q*, Q's causal powers *derive from*, and are *determined by* and *dependent on*, those of Q*.

It is interesting to note that, unlike *Seepage*, none of these alternatives seem to be vulnerable to the drainage argument. The reason is that these alternatives, insofar as we understand them, don't appear to have eliminative implications for causation at the higher, supervenient levels. For example, the fact that Q's causal powers are "explained" by the causal powers of its underlying base Q* does not mean that the former are in any sense preempted or eliminated by the latter, or even that they are somehow reduced to the latter. Exactly what "constitution"[30] might mean, or what "derivation" and "dependence" amount to, requires further thought, but it is clear that these

30. For a defense of nonreductive physicalism based on the idea of constitution, see Derk Pereboom, "Robust Nonreductive Physicalism," *Journal of Philosophy* 99 (2002): 499–531. I believe that the main burden, which is yet to be discharged, of this approach is to produce a serviceably clear concept of constitution. See also Lynne Rudder Baker, *Persons and Bodies: A Constitution View* (Cambridge: Cambridge University Press, 2000).

terms as understood in their rough ordinary philosophical senses have no obviously eliminative intimations.

So why not embrace one or another, or perhaps a combination, of these alternative ways of conceiving the interlevel causal relationships? That would stop the drainage right at the start, and whether there is, or is not, a bottom level makes no difference. So why not say that M, though it doesn't quite have the causal status of P in relation to P*, is a "derivative" cause of P* in virtue of its supervenience on P? M is not in itself an independent cause of P*; its causal status derives from its supervenience on the causally active P. Some years back, I thought that this might be a plausible way of vindicating mental causation.[31] This was the model of so-called supervenient causation. But it soon began to dawn on me that this was an empty verbal ploy; we can "say," if we want, that M is a "supervenient" cause, "dependent" or "derivative" cause, or whatever, and we can embellish *figure 1* by drawing a horizontal arrow connecting M with M*, with the annotation "superveniently causes," as in figure 3. But this is only a gimmick with no meaning; the facts are as represented in the unadorned figure 1, and inserting a dotted arrow and calling it "supervenient" causation, or anything else (how about "pretend" or "faux" causation), does not alter the situation one bit. It neither adds any new facts nor reveals any hitherto unnoticed relationships. Inserting the extra arrow is not only pointless; it could also be philosophically pernicious if it should mislead us into thinking that we have thereby conferred on M, the mental event, some real causal role. Moreover, embracing this approach would lead us back to the overdetermination/exclusion problem—unless we simply stipulate the problem away by declaring that supervenient causal relations do not compete with the causal relation underlying them.

31. In, e.g., "Epiphenomenal and Supervenient Causation," *Midwest Studies in Philosophy* 9 (1984): 257–270. See also Ernest Sosa, "Mind-Body Interaction and Supervenient Causation," *Midwest Studies in Philosophy* 9 (1984): 271–81.

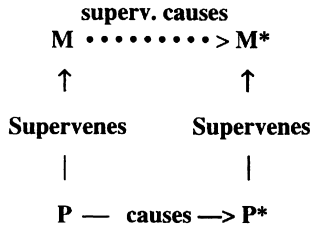

Figure 3.

Jonathan Edwards would have approved my position: in his argument against persisting objects, he did not settle for derivative or dependent causation between created substances; he felt, rightly, that God wholly preempted causation at "higher" levels.

What is Block's own position in regard to these issues? He writes:

> But there is another point of view that recognizes causal efficacy at many levels and does not regard them as competing. And this latter point of view also avoids the problem of causal powers draining away.[32]

And, concerning the "tension" I described at steps (1) and (2) of the supervenience argument, Block writes: "Of course, the non-reductive materialist who accepts causation at many levels should not recognize any tension."[33]

Block's position is the favored approach of most nonreductive physicalists; however, this popular position is precisely what is being challenged. The nonreductive physicalist who accepts supervenience *ought to* recognize the tension; in view of

32. Block, "Do Causal Powers Drain Away?" p. 149. Terence Horgan, among others, holds a similar view; see his "Nonreductive Physicalism and the Explanatory Autonomy of Psychology," in *Naturalism: A Critical Appraisal*, ed. Stephen J. Wagner and Richard Warner (Notre Dame: IN: Notre Dame University Press, 1993).

33. Block, "Do Causal Powers Drain Away?" p. 135.

the considerations advanced in the supervenience argument—basically, Jonathan Edwards's insight—I believe the nonreductive physicalist owes us an explanation of why there is no tension here. It would be nice if we could embrace causation at many levels, including the psychological, the biological, and so on, and also cross-level causation, both downward and upward, all of them coexisting in harmony. And it *is* important to us to be able to have trust in the causal efficacy of our beliefs and desires, emotions and consciousness, and to believe in our powers as agents in the world—all this without reducing mentality to mere patterns of electrical activity in the brain. But these are only a wish list—the starting point of the mental causation debate. The main purpose of the supervenience argument is to bring into focus the disquieting fact that there are strong metaphysical pressures on our pre-philosophical assumptions and desiderata in this area. If the argument is correct, it shows that there are inevitable causal entanglements between different levels, raising all sorts of issues concerning causal closure, competition, and exclusion, and forcing some significant philosophical choices. The nonreductive materialist must sort out and come to terms with these issues; ignoring them is not an option for him. With his drainage argument, Block attempts to defeat the supervenience argument. That is a first step. But this argument has the form of a reductio: if it works, we will know the argument cannot be sound, but that will not tell us just where the argument goes wrong. And this knowledge is required if we are to construct a positive account of multilevel causation of the sort that Block and others have in mind.

In any case, what can be said to counter the drainage argument as lately formulated? As far as the dialectics of the mental causation debate goes, my response here is the same as my reply to Block's statement that the supervenience argument is a reductio against its first premise ("Mental property M causes mental property M*"). As may be recalled, I pointed out that there is another premise against which a reductio can be

performed, namely the premise of psychophysical irreducibility, and that this was my real target. If, as Block's argument suggests, the supervenience argument can be continued to yield as a further conclusion the following proposition:

(H) If there is no bottom level in microphysics, there is no causation anywhere.

and if we find (H) unacceptable, that only means that we need to consider which of the premises of the argument is to be rejected. My suggestion again is that the irreducibility premise should be the prime candidate for rejection; I will elaborate on this below.

Before we go on, there is one point that needs to be clarified. Contrary to what seems sometimes assumed, it is not the case that according to my argument, causation at any level L gives way to causation at level L − 1 (the next lower level), like the rungs of a ladder that keep collapsing each on top of the next lower one. That this is not the case is seen from the fact that the argument requires *Closure* as a premise—the assumption that the lower level in play is causally closed. This means that the mental rung will not collapse onto the biological rung, as far as the supervenience argument is concerned, for the simple reason that the biological level is not causally closed. The same is true of macrolevel physics and chemistry. It is only when we reach the fundamental level of microphysics that we are likely to get a causally closed domain.[34] As I understand it,

34. Actually various complications arise with the talk of levels in this context. In the only levels scheme that has been worked out with some precision, the hierarchical scheme of Paul Oppenheim and Hilary Putnam (in their "Unity of Science as a Working Hypothesis," *Minnesota Studies in the Philosophy of Science*, vol. 2, Minneapolis: University of Minnesota Press, 1958), it is required that each level includes all mereological aggregates of entities at that level (that is, each level is closed under mereological summation). Thus, the bottom level of elementary particles, in this scheme, is in effect the universal domain that includes molecules, organisms, and the rest.

the so-called Standard Model is currently taken to represent the bottom level. Assume that this level is causally closed; the supervenience argument, if it works, shows that mental causal relations give way to causal relations at this microlevel. And similarly for biological causation, chemical causation, geological causation, and the rest. So as far as the supervenience argument goes, the bottom level of fundamental particles (assuming that this is the only level that is causally closed) is always the reference physical domain; there is no step-by-step devolution of causal relations from level to level (I am not suggesting that Block thinks that).[35]

Block's drainage argument evokes some deep metaphysical associations, and this is part of what makes it so interesting. Just think of the whole group of celebrated philosophical arguments with a similar structure, going back to Aristotle and Aquinas. I have in mind Aristotle's argument for the existence of a "prime mover"—the unmoved mover that is the source of all motion. If something moves, it is moved by another thing that moves, which in turn is moved by yet another mover, and so on; but this series cannot go on ad infinitum, for that would make motion impossible. So there must be a mover that is itself not moved by anything else. Aquinas's cosmological argument for the existence of God appears to work in a similar way: there must be a first cause that is itself uncaused because the causal series cannot extend into the past without end. If it did, nothing would exist. The classic foundationalist argument, such as we find in Chisholm,[36] for the existence of "basic" knowledge runs the same way, as do the familiar arguments for

35. A similar problem may well arise for mind-body supervenience; it is likely that mental properties do not supervene on biological properties alone, and that to get full supervenience we have to reach further down and include nonbiological physicochemical properties in the base.

36. Roderick M. Chisholm, *Theory of Knowledge*, 2nd edition (Englewood Cliffs, NJ: Prentice-Hall, 1977).

the existence of semantic primitives, the existence of intrinsic goods, and the like. I think it would be interesting to analyze the metaphysics and logic of arguments that share this general structure. Here, however, I will only make a couple of points specifically in regard to Block's drainage argument.

The first point concerns causal closure. As earlier noted, a causal collapse to the level below would occur only if the lower level is causally closed. Are we assuming that if matter is infinitely divisible, physics will be causally closed at each level of decomposition? I believe that the physicist David Bohm made the observation that each time we descend to a lower microlevel, we do so because the current level is not causally closed ("explanatorily complete" may be a better term in this context); that is, because there are phenomena at this level that can only be explained by descending to a lower level. If something like that is true, no level in Block's infinitely descending series of levels will be causally closed, or explanatorily complete, and the supervenience argument cannot get a toehold. We would not have the required closure premise available—unless we take as our lowest level the *union* of all the microlevels in this infinite chain. Will such a union be causally closed? It has to be, and I believe it may well give us the bottom level which will stop Block's infinite causal drainage.

Second, we must return to reduction again. For Block's drainage argument to work in full force, it must be assumed that the irreducibility premise will hold for purely physical levels—we must assume that molecular facts are not reducible to atomic facts, that atomic facts are not reducible to facts at the level of the Standard Model, and so on down the line. How plausible is this assumption? There are well-known, though by no means undisputed, arguments for regarding the mental to be physically irreducible, and arguments have been advanced to show that the biological level is irreducible to the physicochemical level. But I know of no argument, other than Block's multiple-composition argument discussed above, to

show that the irreducibility assumption will stand as we go down from one microphysical level to the next. The standard view, as I understand it, is that chemistry and macrophysics are reducible, and in fact have already been substantially reduced, to particle physics via quantum mechanics.[37] Unless we have reason to think that irreducibility will hold "all the way down," we have no reason to think that the causal drainage will go on forever. Reduction is the stopper that will plug the cosmic hole through which causal powers might drain away.

In fact, there appear to be presumptive reasons for thinking that reducibility will hold for the kind of infinite series Block has in mind. Let us begin by noting that in various philosophical contexts the identity "the property of being water = the property of being H_2O" is often affirmed and accepted. This identity is accepted presumably on the basis of the fact that water = H_2O. Let us think a bit about what is involved. The property of being H_2O is a total micro-based property of water at the atomic/molecular level; it is the property of being made up of two hydrogen atoms and one oxygen atom in a certain relational structure. Being water is having this kind of microstructure. Having this microstructure is the microstructural essence of water, and being water just is having that structure. We must expect this line of thought to generalize downward, and the following may be one way to flesh it out. Let us say that the property of being H_2O is the total micro-based property of water at the atomic level L (so having M_L = being H_2O). So we have:

(1) Being water = having M_L.

At the next level down, L−1, say the level of the Standard Model, hydrogen atoms have a certain microstructural composition as do oxygen atoms, and water has a certain microstructural

37. See, e.g., Brian P. McLaughlin, "The Rise and Fall of British Emergentism," in *Emergence or Reduction?* ed. Ansgar Beckermann, Hans Flohr, and Jaegwon Kim (Berlin: De Gruyter, 1992).

composition at this level; call it M_{L-1}. Then by the same reasoning that led us to (1), we have:

(2) Being water = having M_{L-1}.

At the level L−2, the one below the Standard Model (if there is such a level), water is again going to have a certain microstructure at that level; this is M_{L-2}. We then have:

(3) Being water = having M_{L-2}.

and so on down the line, to M_{L-3} and the rest. These identities in turn imply the following series of identities:

$M_L = M_{L-1} = M_{L-2} = M_{L-3} \ldots .$

Voilà! These are the identities we need to stop the drainage.

The foregoing is somewhat sketchy and perhaps too quick, and I do not wish to rest my reply to Block's drainage challenge wholly on these rather speculative thoughts. The primary response to the drainage argument is the point that for downward causal drainage to occur, the reduction option must be ruled out for purely physical levels, including microphysical levels, and it is far from obvious that this can be done. In fact, the drainage problem provides us with one more reason to perform a reductio against the irreducibility premise of the supervenience/exclusion argument.