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).