PHILOSOPHY OF MIND

THIRD EDITION

JAEGWON KIM

A Member of the Perseus Books Group

Westview Press was founded in 1975 in Boulder, Colorado, by notable publisher and intellectual Fred Praeger. Westview Press continues to publish scholarly titles and high-quality undergraduate- and graduate-level textbooks in core social science disciplines. With books developed, written, and edited with the needs of serious nonfiction readers, professors, and students in mind, Westview Press honors its long history of publishing books that matter.

Published by Westview Press,
A Member of the Perseus Books Group

Find us on the World Wide Web at www.westviewpress.com.

Westview Press books are available at special discounts for bulk purchases in the United States by corporations, institutions, and other organizations. For more information, please contact the Special Markets Department at the Perseus Books Group, 2300 Chestnut Street, Suite 200, Philadelphia, PA 19103, or call (800) 810-4145, ext. 5000, or e-mail special.markets@perseusbooks.com.

Designed by Trish Wilkinson
Set in 10.5 point Minion Pro

Library of Congress Cataloging-in-Publication Data

Kim, Jaegwon.
Philosophy of mind / Jaegwon Kim.—3rd ed.
p. cm.
ISBN 978-0-8133-4458-4 (alk. paper)
1. Philosophy of mind. I. Title.
BD418.3.K54 2011
128'.2—dc22
E-book ISBN 978-0-8133-4520-8 2010040944

10 9 8 7 6 5 4 3 2 1

CHAPTER 7

Mental Causation

A memorable illustration of mental causation occurs in a celebrated episode in the beginning pages of Proust's *Remembrance of Things Past*. One cold, dreary winter day, the narrator's mother offers him tea, and he takes it with one of the little cakes, "petites madeleines," soaked in it. Here is what happens:

> No sooner had the warm liquid mixed with the crumbs touched my palate than a shudder ran through me and I stopped, intent upon the extraordinary thing that was happening to me. An exquisite pleasure had invaded my senses, something isolated, detached, with no suggestion of its origin. And at once the vicissitudes of life had become indifferent to me, its disasters innocuous, its brevity illusory—this new sensation having had on me the effect which love has of filling me with a precious essence.

The narrator is puzzled: Where does this sudden sense of bliss and contentment come from? Soon, a torrential rush of memories from the distant past is unleashed:

> And suddenly the memory revealed itself. The taste was that of the little piece of madeleine which on Sunday mornings at Combray (because on those mornings I did not go out before mass), when I went to say good morning to her in her bedroom, my aunt Léonie used to give me, dipping it first in her own cup of tea or tisane. . . .
>
> And as soon as I had recognized the taste of the piece of madeleine soaked in her decoction of lime-blossom which my aunt used to give me . . . immediately the old grey house upon the street, where her room

> was, rose up like a stage set to attach itself to the little pavilion opening on to the garden which had been built out behind it for my parents; and with the house the town, from morning to night and in all weathers, the Square where I used to be sent before lunch, the streets along which I used to run errands, the country roads we took when it was fine. And as in the game wherein the Japanese amuse themselves by filling a porcelain bowl with water and steeping in it little pieces of paper which until then are without character or form, but, the moment they become wet, stretch and twist and take on colour and distinctive shape, become flowers or houses or people, solid and recognizable, so in that moment all the flowers in our garden and in M. Swann's park, and the water-lilies on the Vivonne and the good folk of the village and their little dwellings and the parish church and the whole of Combray and its surroundings, taking shape and solidity, sprang into being, town and gardens alike, from my cup of tea.[1]

So begins Proust's journey into the past, in "search of lost time," which takes him more than a dozen years to complete, spanning three thousand pages. All this triggered by some madeleine crumbs soaked in a cup of tea.

This is a case of the so-called involuntary memory—where sensory or perceptual cues we encounter cause recollections of past experiences without conscious effort. It is amazing how a whiff of smell, or a tune, can bring back, totally unexpectedly, a rich panorama of images from a distant past that was apparently lost to us forever.

Returning to our philosophical concerns from the enchanting world of Proust's masterpiece, we can see in this episode several cases of causation involving mental events. The most prominent instance, one of wide literary fame, occurs when the taste of tea-soaked madeleine crumbs causes a sudden burst of recollections of the past. This is a case of mental-to-mental causation—a mental event causing another. There is also the madeleine causing our narrator to experience its taste, a case of physical-to-mental causation. The narrator first declines the tea offered by his mother, then changes his mind and takes the tea. This involves mental-to-physical causation.

When we look around, we see mental causation everywhere. In perception, objects and events around us—the computer display I am staring at, the jet passing overhead, the ocean breeze in the morning—cause visual, auditory, tactile, and other sorts of experiences. In voluntary action, our desires and intentions cause our limbs to move so as to rearrange the objects around us. On a grander scale, it is human knowledge, wishes, dreams, greed, passions, and inspirations that led our forebears to build the pyramids of Egypt and the

Great Wall of China, and to create the glorious music, literature, and artworks that form our cultural heritage. These mental capacities and functions have also been responsible for nuclear weapons, global warming, disastrous oil spills, and the destruction of the rain forests. Mental phenomena are intricately and seamlessly woven into the complex mosaic of causal relations of the world. Or so it seems, at least.

If your mind is going to cause your limbs to move, it presumably must first cause an appropriate neural event in your brain. But how is that possible? How can a mind, or a mental phenomenon, cause a bundle of neurons to fire? Through what mechanisms does a mental event, like a thought or a feeling, manage to initiate, or insert itself into, a causal chain of electrochemical neural events? And how is it possible for a chain of physical and biological events and processes to burst, suddenly and magically, into a full-blown conscious experience, with all its vivid colors, shapes, smells, and sounds? Think of your total sensory experience right now—visual, tactual, auditory, olfactory, and the rest: How is it possible for all this to arise out of molecular activities in the gray matter of your brain?

Agency and Mental Causation

An agent is someone with the capacity to *perform actions for reasons*, and most actions involve bodily movements. We are all agents in that sense: We do such things as turning on the stove, heating water in a kettle, making coffee, and entertaining friends. An action is something we "do"; it is unlike a "mere happening," like sweating, running a fever, or being awakened by the noise of a jackhammer. These are what happens to us; they are not in our control. Implicit in the notion of action is the idea that an agent is in control of what she does, and the control here can only mean causal control.

Let us look into this in some detail. Consider Susan's heating water in a kettle. This must at least include her *causing* the water in the kettle to rise in temperature. Why did Susan heat the water? When someone performs an action, it always makes sense to ask why, even if the correct answer may be "For no particular reason." Susan, let us suppose, heated water to make tea. That is, she *wanted* to make tea and *believed* that she needed hot water to do that—and to be boringly detailed, she *believed* that by heating water in the kettle she could get the hot water she needed. When we know all this, we know why Susan heated water; we understand her action. Beliefs and desires guide actions, and by citing appropriate beliefs and desires, we are able to explain and make sense of why people do what they do.[2]

We may consider the following statement as the fundamental principle that connects desire, belief, and action:

> *Desire-Belief-Action Principle.* (DBA): If agent S desires something and believes that doing A is an optimal way of securing it, S will do A.

As stated, DBA is too strong. For one thing, we often choose not to act on our desires, and sometimes we change them, or try to get rid of them, when we realize that pursuing them is too costly and may lead to consequences that we want to avoid. For example, you wake up in the middle of the night and want a glass of milk, but the thought of getting out of bed in the chilly winter night and going down two long flights of stairs to the dark kitchen talks you out of it. Further, even when we are ready to act on our desires and beliefs, we may find ourselves physically unable to perform the action: It may be that when you have finally overcome your aversion to getting out of the bed, you find yourself chained to the bedposts!

To save DBA, we can tinker with it in various ways; for example, we can add further conditions to the antecedent of DBA (such as that there are no other conflicting desires) or weaken the consequent (for example, by turning it into a probability or tendency statement, or adding the all-purpose hedge "other things being equal" or "under normal conditions"). In any event, there seems little question that a principle like DBA is fundamental to the way we explain and understand actions, both our own and those of others around us. DBA is often taken to be the fundamental schema that anchors reason-based explanations of actions, or "rationalizations." In saying this, we need not imply that beliefs and desires are the only possible reasons for actions; for example, emotions and feelings are often invoked as reasons, as witness, "I hit him because he insulted my wife and that made me angry," or, "He jumped up and down for joy."[3]

What the exceptions to DBA we have considered show is that an agent may have a reason—a "good" reason—to do something but fail to do it. Sometimes there may be more than one belief-desire pair that is related to a given action, as specified by DBA: In addition to your desire for a glass of milk, you heard a suspicious noise from downstairs and wanted to check it out. Let us suppose that you finally did get out of bed to venture down the stairway. Why did you do that? What explains it? It is possible that you went downstairs because you thought you really ought to check out the noise, not out of your desire for milk. If so, it is your desire to check the noise, not your desire for milk, that explains why you went downstairs in the middle of the night. It would be

correct for you to say, "I went downstairs because I wanted to check out the noise," but incorrect to say, "I went downstairs because I wanted a glass of milk," although you did get your milk too. We can also put the point this way: Your desire to check out the noise and your desire for milk were both *reasons*—in fact, *good reasons*—for going downstairs, but the first, not the second, was the *reason for which* you did what you did; it was the *motivating reason*. And it is "reason for which," not mere "reason for," that explains the action. But what precisely is the difference between them? That is, what distinguishes explanatory reasons from reasons that do no explanatory work?

A widely accepted—though by no means undisputed—answer defended by Donald Davidson is the simple thesis that a reason for which an action is done is one that *causes* it.[4] That is, what makes a reason for an action an explanatory reason is its role in the causation of that action. Thus, on Davidson's view, the crucial difference between my desire to check out the noise and my desire for a glass of milk lies in the fact that the former, not the latter, caused me to go downstairs. This makes explanation of action by reasons, or "rationalizing" explanation, a species of causal explanation: Reasons explain actions in virtue of being their causes.

If this is correct, it follows that agency is possible only if mental causation is possible. For an agent is someone who is able to act for reasons and whose actions can be explained and evaluated in terms of the reasons for which she acted. This entails that reasons—that is, mental states like beliefs, desires, and emotions—must be able to cause us to do what we do. Since what we do almost always involves movements of our limbs and other bodily parts, this means that agency—at least human agency—presupposes the possibility of mental-to-physical causation. Somehow your beliefs and desires cause your limbs to move in appropriate ways so that in ten seconds you find your whole body, made up of untold billions of molecules and weighing over a hundred pounds, displaced from your bedroom to the kitchen. A world in which mental causation does not exist is one in which there are no agents and no actions.

Mental Causation, Mental Realism, and Epiphenomenalism

Perception involves the causation of mental events—perceptual experiences and beliefs—by physical processes. In fact, the very idea of perceiving something—say, seeing a tree—involves the idea that the object seen is a cause of your visual experience. Suppose that there is a tree in front of you and that you are having a visual experience of the sort you would be having if your retinas were stimulated

by the light rays reflected by the tree. But you would not be seeing the tree if a holographic image of a tree, visually indistinguishable from the tree, were interposed between you and the tree. You would be seeing the holographic image of a tree, not the tree, even though your perceptual experience in the two cases would have been exactly alike. Evidently, this difference too is a causal one: Your visual experience is caused by a tree holograph, not by the tree.

Perception is our sole window on the world; without it, we could learn nothing about what goes on around us. If, therefore, perception necessarily involves mental causation, there could be no knowledge of the world without mental causation. Moreover, a significant part of our knowledge of the world is based on experimentation, not mere observation. Experimentation differs from passive observation in that it requires our active intervention in the course of natural events; we design and deliberately set up the experimental conditions and then observe the outcome. This means that experimentation presupposes mental-to-physical causation and is impossible without it. Much of our knowledge of causal relations—in general, knowledge of what happens under what conditions—is based on experimentation, and such knowledge is essential not only to our theoretical understanding of the world but also to our ability to predict and control the course of natural events. We must conclude, then, that if minds were not able to causally connect with physical events and processes, we could have neither the practical knowledge required to inform our decisions and actions nor the theoretical knowledge that gives us an understanding of the world around us.

Mental-to-mental causation also seems essential to human knowledge. Consider the process of inferring one proposition from another. Suppose someone asks you, "Is the number of planets odd or even?" If you are like most people, you would probably proceed like this: "Well, how many planets are there? Eight, of course, and eight is an even number because it is a multiple of two. So the answer is: The number is even." You have just inferred the proposition that there are an even number of planets from the proposition that there are eight planets, and you have formed a new belief based on this inference. This process evidently involves mental causation: Your belief that the number of planets is even was caused, through a chain of inference, by your belief that there are eight planets. Inference is one way in which beliefs generate other beliefs. A brief reflection makes it evident that most of our beliefs are generated by other beliefs we hold, and "generation" here could only mean causal generation. It follows, then, that all three types of mental causation—mental-to-physical, physical-to-mental, and mental-to-mental—are implicated in the possibility of human knowledge.

Epiphenomenalism is the view that although all mental events are caused by physical events, they are only "epiphenomena"—that is, events without powers to cause any other event. Mental events are effects of physical (presumably neural) processes, but they do not in turn cause anything else, being powerless to affect physical events or even other mental events; they are the absolute termini of causal chains. The noted nineteenth-century biologist T. H. Huxley has this to say about the consciousness of animals:

> The consciousness of brutes would appear to be related to the mechanism of their body simply as a collateral product of its working and to be as completely without any power of modifying that working as the steam-whistle which accompanies the work of a locomotive engine is without influence upon its machinery. Their volition, if they have any, is an emotion indicative of physical changes, not a cause of such changes.

What about human consciousness? Huxley goes on:

> It is quite true that, to the best of my judgment, the argumentation which applies to brutes holds equally good of men; and, therefore, that all states of consciousness in us, as in them, are immediately caused by molecular changes of the brain-substance. It seems to me that in men, as in brutes, there is no proof that any state of consciousness is the cause of change in the motion of the matter of organism. . . . We are conscious automata.[5]

What was Huxley's argument that convinced him that the consciousness of animals is causally inert? Huxley's reasoning appears to have been something like this: In animal experiments (Huxley mentions experiments with frogs), it can be shown that animals are able to perform complex bodily operations when we have compelling neuroanatomical evidence that they cannot be conscious, and this shows that consciousness is not needed as a cause of these bodily behaviors. Moreover, similar phenomena are observed in cases involving humans: As an example, Huxley cites the case of a brain-injured French sergeant who was reduced to a condition comparable to that of a frog with the anterior part of its brain removed—that is, we have ample anatomical reason to believe that the unfortunate war veteran had no capacity for consciousness—but who could perform complex actions of the kind that we normally think require consciousness, like avoiding obstacles when walking around in a familiar place, eating and drinking, dressing and undressing, and going to bed at the accustomed time. Huxley takes cases of this kind as a basis for his claim that consciousness is not a

cause of behavior production in animals or humans. Whether Huxley's reasoning is sound is something to think about.[6]

Consider a moving car and the series of shadows it casts as it races along the highway: The shadows are caused by the moving car but have no effect on the car's motion. Nor are the shadows at different times causally connected: The shadow at a given instant *t* is caused not by the shadow an instant earlier but by the car itself at *t*. A person who observes the moving shadows but not the car may be led to attribute causal relations between the shadows, the earlier ones causing the later ones, but he would be mistaken. Similarly, you may think that your headache has caused your desire to take aspirin, but that, according to the epiphenomenalist, would be a similar mistake: The headache and the desire for aspirin are both caused by two successive states of the brain, but they are not related as cause to effect any more than two successive shadows of the moving car. The apparent regularities that we observe in mental events, the epiphenomenalist argues, do not represent genuine causal connections; like the regularities characterizing the car's moving shadows or the successive symptoms of a disease, they are merely reflections of the real causal processes at a more fundamental level.

These are the claims of epiphenomenalism. Few philosophers have been self-professed epiphenomenalists, although there are those whose views appear to lead to such a position (as we will see below). We are more likely to find epiphenomenalist thinking among scientists in brain science. At least, some scientists seem to treat mentality, especially consciousness, as a mere shadow or afterglow thrown off by the complex neural processes going on in the brain; these physical-biological processes are what at bottom do all the pushing and pulling to keep the human organism functioning. If conscious events really had causal powers to influence neural events, there could be no complete neural-physical explanations of neural events unless consciousness was explicitly brought into neuroscience as an independent causal agent in its own right. That is, there could be no complete physical-biological theory of neural phenomena. It would seem that few neuroscientists would countenance such a possibility. (For further discussion, see chapter 10.)

How should we respond to the epiphenomenalist stance on the status of mind? Samuel Alexander, a leading emergentist during the early twentieth century, comments on epiphenomenalism with a pithy remark:

> [Epiphenomenalism] supposes something to exist in nature which has nothing to do, no purpose to serve, a species of noblesse which depends on the work of its inferiors, but is kept for show and might as well, and undoubtedly would in time, be abolished.[7]

Alexander is saying that if epiphenomenalism is true, mind has no work to do and hence is entirely useless, and it is pointless to recognize it as something real. Our beliefs and desires would have no role in causing our decisions and actions and would be entirely useless in their explanations; our perception and knowledge would have nothing to do with our artistic creations or technological inventions. *Being real and having causal powers go hand in hand; to deprive the mind of causal potency is in effect to deprive it of its reality.*

It is important to see that this is not an *argument* against epiphenomenalism: Alexander only points out, in a stark and forceful way, what accepting epiphenomenalism would entail. We should also remind ourselves that the typical epiphenomenalist does not reject the reality of mental causation altogether; she only denies mind-to-body and mind-to-mind causation, not body-to-mind causation. In this sense, she gives the mental a well-defined place in the causal structure of the world; mental events are integrated into that structure as effects of neural processes. This suggests that there is a stronger form of epiphenomenalism, according to which the mental is both causeless and effectless—that is, the mental is simply noncausal. To a person holding such a view, mental events are in total causal isolation from the rest of the world, even from other mental events; each mental event is a solitary island, with no connection to anything else. (Recall the discussion of the causal status of immaterial substances, in chapter 2.) Its existence would be entirely inexplicable since it has no cause, and it would make no difference to anything else since it has no effect. It would be a mystery how the existence of such things could be known to us. As Alexander declares, they could just as well be "abolished"—that is, regarded as nonexistent. No philosopher appears to have explicitly held or argued for this stronger form of epiphenomenalism; however, as we will see, there are views on the mind-body problem that seem to lead to a radical epiphenomenalism of this kind.

So why not grant the mind full causal powers, among them the power to influence bodily processes? This would give the mental a full measure of reality and recognize what after all is so manifestly evident to common sense. That is just what Descartes tried to do with his thesis that minds and bodies, even though they are substances of very different sorts, are in intimate causal commerce with each other. But we have seen what seem like impossible difficulties besetting his program (chapter 2).

Everyone will acknowledge that mental causation is a desideratum—something important to save. Jerry Fodor is not jesting when he writes:

> I'm not really convinced that it matters very much whether the mental is physical; still less that it matters very much whether we can prove that it is.

> Whereas, if it isn't literally true that my wanting is causally responsible for my reaching, and my itching is causally responsible for my scratching, and my believing is causally responsible for my saying . . . , if none of that is literally true, then practically everything I believe about anything is false and it's the end of the world.[8]

For Fodor, then, mental causation is absolutely nonnegotiable. And it is understandable why anyone should feel this way: Giving up mental causation amounts to giving up our conception of ourselves as agents and cognizers. Is it even *possible* for us to give up the idea that we are agents who decide and act, that we perceive and know certain things about the world? Can we live our lives as epiphenomenalists? That is, as "practicing" epiphenomenalists?

In his first sentence in the foregoing quoted passage, Fodor is saying that being able to defend a theory of the mind-body relation is far less important than safeguarding mental causation. That is not an implausible perspective to take: Whether a stance on the mind-body problem is acceptable depends importantly, if not solely, on how successful it is in giving an account of mental causation. On this criterion, Descartes's substance dualism, in the opinion of many, must be deemed a failure. So the main question is this: Which positions on the mind-body problem allow full-fledged mental causation and provide an explanation of how it is possible? We consider this question in the sections to follow.

Psychophysical Laws and "Anomalous Monism"

The expulsion of Cartesian immaterial minds perhaps brightens the prospect of understanding how mental causation is possible. For we no longer have to contend with a seemingly hopeless question: How could immaterial souls with no physical characteristics—no bulk, no mass, no energy, no charge, and no location in space—causally influence, and be influenced by, physical objects and processes? Today few, though not all, philosophers or scientists regard minds as substances of a special nonphysical sort; mental events and processes are now viewed as occurring in complex physical systems like biological organisms, not in immaterial minds. The problem of mental causation, therefore, is now formulated in terms of two kinds of events, mental and physical, not in terms of two kinds of substances: How is it possible for a mental event (such as a pain or a thought) to cause a physical event (a limb withdrawal, an utterance)? Or in terms of properties: How is it possible for an

instantiation of a mental property (for example, that of experiencing pain) to cause a physical property to be instantiated?

But why is this supposed to be a "problem"? We do not usually think that there is a special philosophical problem about, say, chemical events causally influencing biological processes or a nation's economic and political conditions causally affecting each other. So what is it about mentality and physicality that make causal relations between them a philosophical problem? For substance dualism, it is, at bottom, the extreme heterogeneity of minds and bodies, in particular, the nonspatiality of minds and the spatiality of bodies (as argued in chapter 2), that makes causal relations between them problematic. Given that mental substances have now been expunged, aren't we home free with mental causation? The answer is that certain other assumptions and doctrines that demand our respect present apparent obstacles to mental causation.

One such doctrine centers on the question of whether there are *laws connecting mental phenomena with physical phenomena*—that is, psychophysical laws—that are thought to be needed to underwrite causal connections between them. Donald Davidson's well-known "anomalism of the mental" states that there can be no such laws.[9] A principle connecting laws and causation that is widely, though not universally, accepted, is this: *Causally connected events must instantiate, or be subsumed under, a law*. If heating a metallic rod causes its length to increase, there must be a law connecting events of the first type and events of the second type; that is, there must be a law stating that the heating of a metallic rod is followed by an increase in its length. But if causal connections require laws and there are no laws connecting mental events with physical events, it would seem to follow that there could be no mental-physical causation. This line of reasoning is examined in more detail later in the chapter. But is there any reason to doubt the existence of laws connecting mental and physical phenomena?

In earlier chapters, we often assumed that there are lawful connections between mental and physical events; you surely recall the stock example of pain and C-fiber excitation. The psychoneural identity theory, as we saw, assumes that each type of mental event is lawfully correlated with a type of physical event. Talk of "physical realization" of mental events also presupposes that there are lawlike connections between a mental event of a given kind and its diverse physical realizers, for a physical realizer of a mental event must at least be sufficient, as a matter of law, for the occurrence of that mental event. The very idea of a "neural correlate" seems to imply that there are psychophysical laws; if a mental state and its neural correlate co-occur, that has to be a lawlike relationship, not an accidental connection. Davidson explicitly restricts his

claim about the nonexistence of psychophysical laws to intentional mental events and states ("propositional attitudes")—that is, states with propositional content, like beliefs, desires, hopes, and intentions; he is not concerned with sensory events and states, like pains, visual sensings of color, and mental images. Why does Davidson think that no laws exist connecting, say, beliefs with physical-neural events? Doesn't every mental event have a neural substrate, that is, a neural state that, as a matter of law, suffices for its occurrence?

Before we take a look at Davidson's argument, let us consider some examples. Take the belief that it is unseemly for the president of the United States to get a $500 haircut. How reasonable is it to expect to find a neural substrate for this belief? Is it at all plausible to think that all and only people who have this belief share some specific neural state? It makes perfectly good sense to try to find neural correlates for pains, sensations of thirst and hunger, visual images, and the like, but somehow it does not seem to make much sense to look for the neural correlates of mental states like our sample belief, or for things like your sudden realization that you have a philosophy paper due in two days, your hope that airfares to California will come down after Christmas, and the like. Is it just that these mental states are so complex that it is very difficult, perhaps impossible, for us to discover their neural bases? Or is it the case that they are simply not the sort of state for which neural correlates could exist and that it makes no sense to look for them?

This is not intended as an argument for the impossibility of psychophysical laws but it should dispel, or at least weaken, the strong presumption many of us are apt to hold that there must "obviously" be psychophysical laws since mentality depends on what goes on in the brain. It is now time to turn to Davidson's famous but notoriously difficult argument against psychophysical laws.[10]

A crucial premise of Davidson's argument is the thesis that the ascription of intentional states, like beliefs and desires, is regulated by certain *principles of rationality* that ensure that the total set of such states attributed to a person will be as rational and coherent as possible. This is why, for example, we refrain from attributing to a person manifestly contradictory beliefs, even when the sentences uttered have the surface logical form of a contradiction. When someone replies, "Well, I do and I don't," when asked, "Do you like Ralph Nader?" we do not take her to be expressing a literally contradictory belief—the belief that she both likes and does not like Nader. Rather, we take her to be saying something like, "I like some aspects of Nader (say, his concerns for social and economic justice), but I don't like other aspects (say, his presidential ambitions)." If she were to insist, "No, I don't mean that; I really both do and don't like Nader, period," we would not know what to make of her; perhaps

her "and" does not mean what the English "and" means, or perhaps she does not have a full grasp of "not." We cast about for some consistent interpretation of her meaning because an interpreter of a person's speech and mental states is under the mandate that an acceptable interpretation must make her come out with a reasonably coherent set of beliefs—as coherent and rational as evidence permits. When no minimally coherent interpretation is possible, we are apt to blame our own interpretive efforts rather than accuse our subject of harboring explicitly inconsistent beliefs. We also attribute to a subject beliefs that are obvious logical consequences of beliefs already attributed to him. For example, if we have ascribed to a person the belief that Boston is less than sixty miles from Providence, we would, and should, ascribe to him the belief that Boston is less than seventy miles from Providence, the belief that Boston is less than one hundred miles from Providence, and countless others. We do not need independent evidence for these further belief attributions; if we are not prepared to attribute any one of these further beliefs, we should reconsider our original belief attribution and be prepared to withdraw it. Our concept of belief does not allow us to say that someone believes that Boston is within sixty miles of Providence but does not believe that it is within seventy miles—unless we are able to give an intelligible explanation of how this could happen in this particular case. This principle, which requires that the set of beliefs be "closed" under obvious logical entailment, goes beyond the simple requirement of consistency in a person's belief system; it requires that the belief system be coherent as a whole—it must in some sense hang together, without unexplainable gaps. In any case, Davidson's thesis is that the requirement of rationality and coherence[11] is of the essence of the mental—that is, it is *constitutive* of the mental in the sense that it is exactly what makes the mental mental. Keep in mind that Davidson is speaking only of intentional states, like belief and desire, not sensory states and events like pains and afterimages. (For further discussion, see chapter 8 on interpretation theory.)

But it is clear that the physical domain is subject to no such requirement; as Davidson says, the principle of rationality and coherence finds "no echo" in physical theory. Suppose now that we have laws connecting beliefs with brain states; in particular, suppose we have laws that specify a neural substrate for each of our beliefs—a series of laws of the form "N occurs to a person at *t* if and only if B occurs to that person at *t*," where N is a neural state and B is a belief. If such laws were available, we could attribute beliefs to a subject, *one by one*, independently of the constraints of the rationality principle. For in order to determine whether she has a certain belief B, all we would need to do is ascertain whether B's neural substrate N is present in her; there would be no

need to check whether this belief makes sense in the context of her other beliefs or even what other beliefs she has. In short, we could read her mind by reading her brain. The upshot is that the practice of belief attribution would no longer be regulated by the rationality principle. By being connected by law with neural state N, belief B becomes hostage to the constraints of physical theory. On Davidson's view, as we saw, the rationality principle is constitutive of mentality, and beliefs that have escaped its jurisdiction can no longer be considered beliefs. If, therefore, belief is to retain its identity and integrity as a mental phenomenon, its attribution must be regulated by the rationality principle and hence cannot be connected by law to a physical substrate.

Let us assume that Davidson has made a plausible case for the impossibility of psychophysical laws (we may call his thesis "psychophysical anomalism") so that it is worthwhile to explore its consequences. One question that was raised earlier is whether it might make mental causation impossible. Here the argument could go like this: Causal relations require laws, and this means that causal relations between mental events and physical events require psychophysical laws, laws connecting mental and physical events. But Davidson's psychophysical anomalism holds that there can be no such laws, whence it would appear to follow that there can be no causal relations between mental and physical phenomena. Davidson, however, is a believer in mental causation; he explicitly holds that mental events sometimes cause, and are caused by, physical events. This means that Davidson must reject the argument just sketched that attempts to derive the nonexistence of mental causation from the nonexistence of psychophysical laws. How can he do that?

What Davidson disputes in this argument is its first step, namely, the inference from the premise that *causation requires laws* to the conclusion that *psychophysical causation requires psychophysical laws*. Let us look into this in some detail. To begin, what is it for one individual event *c* to cause another individual event *e*? This holds, on Davidson's view, only if the two events instantiate a law, in the following sense: *c* falls under a certain event kind (or description) F, *e* falls under an event kind G, and there is a law connecting events of kind F with events of kind G (as cause to effect). This is a form of the influential nomological account of causation: Causal connections must instantiate, or be subsumed under, general laws. Suppose, then, that a particular mental event, *m*, causes a physical event, *p*. This means, according to the nomological conception of causation, that for some event kinds, C and E, *m* falls under C and *p* falls under E, and there is a law that connects events of kind C with events of kind E. This makes it evident that laws connect individual events only as they fall under kinds. Thus, when psychophysical anomal-

ism says that there are no psychophysical laws, what it says is that there are no laws connecting mental kinds with physical kinds. So what follows is only that *if mental event* m *causes physical event* p, *kinds C and E, under which* m *and* p, *respectively*, *fall and which are connected by law, must both be physical kinds*. That is to say, a purely physical law must underwrite this causal relation. In particular, this means that C, under which mental event *m* falls, cannot be a mental kind; it must be a physical one. From which it follows that *m* is a physical event! For an event is mental or physical according to whether it falls under a mental kind or a physical kind. Note that this "or" is not exclusive; *m*, being a mental event, must fall under a mental kind, but that does not prevent it from falling under a physical kind as well. This argument applies to all mental events that are causally related to physical events, and there appears to be no reason not to think that every mental event has some causal connection, directly or via a chain of other events, with a physical event. All such events, on Davidson's argument, are physical events.[12]

That is Davidson's "anomalous monism." It is a monism because it claims that all individual events, mental events included, are physical events (you will recall this as "token physicalism"; see chapter 1). Moreover, it is physical monism that does not require psychophysical laws; in fact, as we just saw, the argument for it requires the nonexistence of such laws, whence the term "anomalous" monism. Davidson's world, then, looks like this: It consists exclusively of physical objects and physical events, but some physical events fall under mental kinds (or have mental descriptions) and therefore are mental events. Laws connect physical kinds and properties with other physical kinds and properties, and these laws generate causal relations between individual events. Thus, all causal relations of this world are exclusively grounded in physical laws.

Is Anomalous Monism a Form of Epiphenomenalism?

One of the premises from which Davidson derives anomalous monism is the claim that mental events can be, and sometimes are, causes and effects of physical events. On anomalous monism, however, to say that a mental event *m* is a cause of an event *p* (*p* may be mental or physical) amounts only to this: *m* has a physical property Q (or falls under a physical kind Q) such that an appropriate law connects Q (or events with property Q) with some physical property P of *p*. Since no laws exist that connect mental and physical properties, purely physical laws must do all the causal work, and this means that individual events can enter into causal relations only because they possess physical properties that

figure in laws. Consider an example: Your desire for a drink of water causes you to turn on the tap. On Davidson's nomological conception of causation, this requires a law that subsumes the two events, your desiring a drink of water and your turning on the tap. However, psychophysical anomalism says that this law must be a physical law, since there are no laws connecting mental kinds with physical kinds. Hence, your desire for a drink of water must be redescribed physically—that is, a suitable physical property of your desire must be identified—before it can be brought under a law. In the absence of psychophysical laws, therefore, it is the physical properties of mental events that determine, wholly and exclusively, what causal relations they enter into. In particular, the fact that your desire for a drink of water is a desire for a drink of water—that is, the fact that it is an event of this mental kind—apparently has no bearing on its causation of your turning on the tap. What is causally relevant is its physical properties—presumably the fact that it is a neural, or physicochemical, event of a certain kind.

It seems, then, that under anomalous monism, mental properties are causal idlers with no work to do. To be sure, anomalous monism is not epiphenomenalism in the classic sense, since individual mental events are allowed to be causes of other events. The point, though, is that it is an epiphenomenalism of *mental properties*—we may call it "mental property epiphenomenalism"[13]—in that it renders mental properties and kinds causally irrelevant. Moreover, it is a form of radical epiphenomenalism described earlier: Mental properties play no role in making mental events either causes or effects. To make this vivid: If you were to redistribute mental properties over the events of this world any way you please—you might even remove them entirely from all events, making all of them purely physical—that would not alter, in the slightest way, the network of causal relations of this world; it would not add or subtract a single causal relation anywhere in the world!

This shows the importance of properties in the debate over mental causation: It is the causal efficacy of mental properties that we need to vindicate and give an account of. With mental substances out of the picture, there are only mental properties left to play any causal role, whether these are construed as properties of events or of objects. If mentality is to do any causal work, it must be the case that having a given mental property rather than another, or having it rather than not having it, must make a causal difference; it must be the case that an event, because it has a certain mental property (for example, being a desire for a drink of water), enters into a causal relation (it causes you to look for a water fountain) that it would otherwise not have entered into. We must

therefore conclude that Davidson's anomalous monism fails to pass the test of mental causation; by failing to account for the causal efficacy and relevance of mental properties, it fails to account for the possibility of mental causation.

The challenge posed by Davidson's psychophysical anomalism, therefore, is to answer the following question: How can anomalous mental properties, properties that are not fit for laws, be causally efficacious properties? It would seem that there are only two ways of responding to this challenge: First, we may try to reject its principal premise, namely, psychophysical anomalism, by finding faults with Davidson's argument and then offering plausible reasons for thinking that there are indeed psychophysical laws that can underwrite psychophysical causal relations. Second, we may try to show that the nomological conception of causality—in particular, as it is construed by Davidson—is not the only way to understand causation and that there are alternative conceptions of causation on which mental properties, though anomalous, could still be causally efficacious. Let us explore the second possibility.

Counterfactuals to the Rescue?

There indeed is an alternative approach to causation that on its face does not seem to require laws, and this is the counterfactual account of causation. On this approach, to say that event *c* caused event *e* is to say that if *c* had not occurred, *e* would not have occurred.[14] The thought that a cause is the *sine qua non* condition, or *necessary* condition, of its effect is a similar idea. This approach has much intuitive plausibility. The overturned space heater caused the house fire. What makes it so? Because if the space heater had not overturned, the fire would not have occurred. What is the basis of saying that the accident was caused by a sudden braking on a rain-slick road? Because if the driver had not suddenly stepped on his brake pedal on the wet road, the accident would not have occurred. In such cases we seem to depend on counterfactual ("what if") considerations rather than laws. Especially if you insist on exceptionless "strict" laws, as Davidson does, we obviously are not in possession of such laws to support these perfectly ordinary and familiar causal claims, claims that we regard as well supported.

The situation seems the same when mental events are involved: There is no mystery about why I think that my desire for a drink of water caused me to step into the dark kitchen last night and stumble over the sleeping dog. It's because of the evidently true counterfactual "If I had not wanted a drink of water last night, I would not have gone into the kitchen and stumbled over the

dog." In confidently making these ordinary causal or counterfactual claims, we seem entirely unconcerned about the question of whether there are laws about wanting a glass of water and stumbling over a sleeping dog. Even if we were to reflect on such questions, we would be undeterred by the unlikely possibility that such laws exist or can be found. To summarize, then, the idea is this: We know that mental events, in virtue of their mental properties, can, and sometimes do, cause physical events because we can, and sometimes do, know appropriate mental-physical counterfactuals to be true. Mental causation is possible because such counterfactuals are sometimes true.

The counterfactual account of causation opens up the possibility of explaining mental causation in terms of how mental-physical counterfactuals can be true. To show that there is a special problem about mental causation, you must show that there are problems about these counterfactuals.

So are there special problems about these psychophysical counterfactuals? Do we have an understanding of how such counterfactuals can be true? There are many philosophical puzzles and difficulties surrounding counterfactuals, especially about their "semantics"—that is, conditions under which counterfactuals can be evaluated as true or false. There are two main approaches to counterfactuals: (1) the nomic-derivational approach, and (2) the possible-world approach.[15] On the nomic-derivational approach, the counterfactual conditional "If P were the case, Q would be the case" (where P and Q are propositions) is true just in case the consequent, Q, of the conditional can be logically derived from its antecedent, P, when taken together with laws and statements of conditions holding on the occasion.[16] Consider an example: "If this match had been struck, it would have lighted." This counterfactual is true since its consequent, "The match lighted," can be derived from its antecedent, "The match was struck," in conjunction with the law "Whenever a dry match is struck in the presence of oxygen, it lights," taken together with the auxiliary premises "The match was dry" and "There was oxygen present."

It should be immediately obvious that on this analysis of counterfactuals, the counterfactual account of mental causation does not make the problem of mental causation go away. For the truth of the psychophysical counterfactuals—like "If I had not wanted to check out the strange noise, I would not have gone downstairs," and, "If Jones's C-fibers had been activated, she would have felt pain"—would require laws that would enable the derivation of the physical consequents from their psychological antecedents (or vice versa), and this evidently requires psychophysical laws, laws connecting mental with physical phenomena. On the nomic-derivational approach, therefore, Davidson's problem of psychophysical laws arises again.

Let us consider then the possible-world approach to the truth conditions of counterfactuals. In a simplified form, it says this: The counterfactual "If P were the case, Q would be the case" is true just in case Q is true in the world in which P is true and that, apart from P's being true there, is as much like the actual world as possible. (To put it another way: Q is true in the "closest" P-world.)[17] To ascertain whether this counterfactual is true, we go through the following steps: Since this is a counterfactual, its antecedent, P, is false in the actual world. We must go to a possible world ("world" for short) in which P is true and see whether Q is also true there. But there are many worlds in which P is true—that is, there are many P-worlds—and in some of these Q is true and in others false. So which P-world should we pick in which to check on Q? The answer: Pick the P-world that is the most similar, or the "closest," to the actual world. The counterfactual "If P were true, Q would be true" is true if Q is true in the closest P-world; it is false if Q is false in that world.

Let us see how this works with the counterfactual "If this match had been struck, it would have lighted." In the actual world, the match was not struck; so suppose that the match was struck (this means, go to a world in which the match was struck), but keep other conditions the same as much as possible. Certain other conditions must be altered under the counterfactual supposition that the match was struck: For example, in the actual world the match lay motionless in the matchbox and there was no disturbance in the air in its vicinity, so these conditions have to be changed to keep the world consistent as a whole. However, we need not, and should not, change the fact that the match was dry and the fact that sufficient oxygen was present in the ambient air. So in the world we have picked, the following conditions, among others, obtain: The match was struck, it was dry, and oxygen was present in the vicinity. The counterfactual is true if and only if the match lighted in that world. Did the match light in that world? In asking this question, we are asking which of the following two worlds is closer to the actual world:[18]

W_1: The match was struck; it was dry; oxygen was present; the match lighted.
W_2: The match was struck; it was dry; oxygen was present; the match did not light.

We would judge, it seems, that of the two, W_1 is closer to the actual world, thereby making the counterfactual come out true. But why do we judge this way?

There seems to be only one answer: Because in the actual world there is a lawful regularity to the effect that when a dry match is struck in the presence

of oxygen, it ignites, and this law holds in W_1, but not W_2. That is why W_1 is closer to the actual world than W_2 is. So in judging that this match, which in fact was dry and bathed in oxygen, would have lighted if it had been struck, we seem to be making crucial use of the law just mentioned. If in the actual world dry matches, when struck in the presence of oxygen, seldom or never light, there seems little question that we would go for W_2 as the closer world and judge the counterfactual "If this match had been struck, it would have lighted" to be false. If this is right, the counterfactual model of causation does not entirely free us from laws, as we had hoped; it seems that at least in some cases we must still resort to laws and lawful regularities.

Now consider a psychophysical counterfactual: "If Brian had not wanted to check out the noise, he wouldn't have gone downstairs." Suppose that we take this counterfactual to be true, and on that basis we judge that Brian's desire to check out the noise caused him to go downstairs. Consider the following two worlds:

W_3: Brian didn't want to check out the noise; he didn't go downstairs.
W_4: Brian didn't want to check out the noise; he went downstairs anyway.

If W_4 is closer to the actual world than W_3 is, that would falsify our counterfactual. So why should we think that W_3 is closer than W_4? In the actual world, Brian wanted to check out the noise and went downstairs. As far as these two particular facts are concerned, W_4 evidently is closer to the actual world than W_3 is. So why do we hold W_3 to be closer and hence the counterfactual to be true? The only plausible answer, again, seems to be something like this: We know, or believe, that there are certain lawful regularities and propensities governing Brian's wants, beliefs, and so on, on the one hand, and his behavior patterns, on the other, and that, given the absence of something like a desire to check out a suspicious noise, along with other conditions prevailing at the time, his not going downstairs at that particular time fits these regularities and propensities better than the supposition that he would have gone downstairs at that time. We consider such regularities and propensities, that is, facts about a person's personality, to be reliable and lawlike and commonly appeal to them in assessing counterfactuals of this kind (and also in making predictions and guesses as to how a person will behave), even though we may have only the vaguest idea about the details and lack the ability to articulate them in a precise way.

Again, the relevance of psychophysical laws to mental causation is apparent. Although there is room for further discussion, it is plausible that considerations of lawful regularities governing mental and physical phenomena often seem crucially involved in the evaluation of psychophysical counterfactuals of the sort that can ground causal relations. We need not know the details of such regularities, but we must believe that they exist and know their rough content and shape to be able to evaluate these counterfactuals as true or false. So are we back where we started, with Davidson and his argument for the impossibility of psychophysical laws?

Not exactly, fortunately. The laws involved in evaluating counterfactuals, as is clear from our examples, need not be laws of the kind Davidson has in mind—what he calls "strict" laws. These are exceptionless, explicitly articulated laws that form a closed and comprehensive theory, like the laws of physics. Rather, the laws involved in evaluating these quotidian counterfactuals—indeed, laws on the basis of which causal judgments are made in much of science—are rough-and-ready generalizations tacitly qualified by generous escape clauses ("ceteris paribus," "under normal conditions," "in the absence of interfering forces," and so on) and apparently immune to falsification by isolated negative instances. Laws of this type, sometimes called "ceteris paribus laws," seem to satisfy the usual criteria of lawlikeness: As we saw, they seem to have the power to ground counterfactuals, and their credence is enhanced as we observe more and more positive instances. Their logical form, their verification conditions, and their efficacy in explanations and predictions are not well understood, but it seems beyond question that they are the essential staple that sustains and nourishes our counterfactuals and causal discourse.[19]

Does the recognition that causal relations involving mental events can be supported by these "nonstrict," ceteris paribus laws solve the problem of mental causation? It does enable us to get around the difficulty raised by Davidsonian considerations—at least for now.

We can see, however, that the difficulty has not been fully resolved. For it may well be that these nonstrict laws are possible only if strict laws are possible and that where there are no underlying strict laws that can explain them or otherwise ground them, they remain only rough, fortuitous correlations without the power to support causal claims. It may be that their lawlike appearance is illusory and that this makes them unfit to ground causal relations. More important, as we said, the nature of these ceteris paribus laws is not well understood; though laws of this kind seem in fact used to back up causal claims, we lack a theoretical understanding of how this works.

Physical Causal Closure and the "Exclusion Argument"

Suppose, then, that we have somehow overcome the difficulties arising from the possibility that there are no mental-physical laws capable of supporting mental-physical causal relations. We are still not home free: There is another challenge to mental causation that we must confront, a challenge that is currently considered to be the gravest threat to the possibility of mental causation. The new threat arises from the principle, embraced by most physicalists, that asserts that the physical domain is *causally closed*. What does this mean? Pick any physical event—say, the decay of a uranium atom or the collision of two stars in distant space—and trace its causal ancestry or posterity as far as you would like; the principle of physical causal closure says that this will never take you outside the physical domain. Thus, no causal chain involving a physical event ever crosses the boundary of the physical into the nonphysical: If x is a physical event and y is a cause or effect of x, then y too must be a physical event.

For present purposes, it is convenient to use a somewhat weaker form of causal closure stated as follows:

> *Causal Closure of the Physical Domain*. If a physical event has a cause (occurring) at time t, it has a sufficient physical cause at t.

Notice a few things about this principle. First, it does not flatly say that a physical event can have no nonphysical cause; all it says is that in our search for its cause, we never need to look outside the physical domain. In that sense, the physical domain is causally, and hence explanatorily, self-sufficient and self-contained. Second, it does not say that every physical event has a sufficient physical cause or a physical causal explanation; in this regard, it differs from physical causal determinism, the thesis that every physical event has a sufficient physical cause. Third, the closure principle is consistent with mind-body dualism: So far as it goes, there might be a separate domain of Cartesian immaterial minds. All it requires is that there be no injection of causal influence into the physical world from outside, including Cartesian minds.

Most philosophers appear to find physical causal closure plausible; of course, anyone who considers himself or herself a physicalist of any kind must accept it. If the closure should fail to hold, there would be physical events for whose explanation we would have to look to nonphysical causal agents, like spirits or divine forces outside space-time. That is exactly the situ-

ation depicted in Descartes's interactionist dualism (chapter 2). If closure fails, theoretical physics would be in principle incompletable, a prospect that few physicists would countenance. It seems clear that research programs in physics, and the rest of the physical sciences, presuppose something like the closure principle.

It is worth noting that neither the biological domain nor the psychological domain—in fact, no domain of a special science—is causally closed: There are nonbiological events that cause biological events (for example, radiation causing cells to mutate; a volcanic eruption wiping out a whole species), and we are familiar with cases in which nonpsychological events cause psychological events (for example, purely physical stimuli causing sensations and perceptual experiences). In any case, physical causal closure gives a meaning to the widely shared view that the physical domain is an all-encompassing domain and that physics, which is the science of this domain, is our basic science. Some consider the closure principle an a posteriori truth overwhelmingly supported by the rise of modern physical science;[20] those who consider the very idea of causal interference in the physical world from some immaterial or transcendental forces incoherent might argue that the closure principle is conceptual and a priori. It is also possible to regard the principle primarily as a methodological-regulative principle that guides research and theory-building in the physical sciences.

At any rate, it is easy to see that the physical closure principle directly creates difficulties for mental causation, in particular mental-to-physical causation. Suppose that a mental event, m, causes a physical event, p. The closure principle says that there must also be a physical cause of p—an event, p^*, occurring at the same time as m, that is a sufficient cause of p. This puts us in a dilemma: Either we have to say that $m = p^*$—namely, identify the mental cause with the physical cause as a single event—or else we have to say that p has two distinct causes, m and p^*, that is, it is causally overdetermined. The first horn turns what was supposed to be a case of mental-to-physical causation into an instance of physical-to-physical causation, a result only a reductionist physicalist would welcome. Grasping the second horn of the dilemma would force us to admit that every case of mental-to-physical causation is a case of causal overdetermination, one in which a physical cause, even if the mental cause had not occurred, would have brought about the physical effect. This seems like a bizarre thing to believe, but quite apart from that, it appears to weaken the status of the mental event as a cause of the physical effect. To vindicate m as a full and genuine cause of p, we should be able to show that m can bring about p on its own, without there being a synchronous physical event that also serves as a sufficient cause of p. According to our reasoning, however, every mental event

has a physical partner that would have brought about the effect anyway, even if the mental cause were taken out of play entirely.

This thought can be developed along the following lines. Consider the following constraint:

> *Exclusion Principle*. No event has two or more distinct sufficient causes, all occurring at the same time, unless it is a genuine case of overdetermination.

Genuine overdetermination is illustrated by the "firing squad" example: Multiple bullets hit a person at the same time, and this kills the person, where a single bullet would have sufficed. A house fire is caused by a short circuit and at the same time by a lightning strike. In these cases, two or more independent causal chains converge on a single effect. Given this, the exclusion principle should look obviously, almost trivially, true.

Return now to our case of mental-to-physical causation. We begin with the assumption that there is a case in which a mental event causes a physical event:

(1) m is a cause of p.

As we saw, it follows from (1) and physical causal closure that there is also a physical event p^*, occurring at the same time as m, such that:

(2) p^* is a cause of p.

Let us suppose further that we don't want (1) to collapse into a case of physical-to-physical causation; that is, we want:

(3) $m \neq p^*$.

Suppose we assume further:

(4) This is not a case of overdetermination.

Given the closure and the exclusion principles, these four propositions put us in trouble: According to (1), (2), and (3), p has two distinct causes, m and p^*; since (4) says that this is not a case of overdetermination, the exclusion principle kicks in, saying that either m or p^* must be disqualified as a cause of p. Which one? The answer: p^* stays, m must go. The reason is simple: If we try to retain m,

the closure principle kicks in again and says that there must also be a physical cause of p—and what could this be if not p^*? Obviously, we are back at the same situation: Unless we eliminate m and keep p^*, we would be off to an infinite regress, or treading water forever in the same place. So our conclusion has to be:

(5) Hence, m is not a cause of p, and (1) is false.

The reasoning obviously generalizes to every putative case of mental causation, and it further follows:

(6) Mental events never cause physical events.

This argument is a form of the much-debated "exclusion argument," since it aims to show how a mental cause of a physical event is always excluded by a physical cause.[21] The apparent moral of the argument is that mental-to-physical causation is illusory; it never happens. This is epiphenomenalism, at least with regard to causation of physical events. It does not exclude mental events causing other mental events. But if mental events, like beliefs and intentions, never cause bodily movements, that makes agency plainly impossible, and Fodor has something to worry about: His world might be coming to an end!

That, anyway, is the way the implications of the argument are usually understood. However, that is not the only way to read the moral of the argument: If we are prepared to reject the antiphysicalist assumption (3) by embracing the mind-body identity "$m = p^*$," we can escape the epiphenomenalist consequence of the argument. If $m = p^*$, here there is only one event and hence only one cause of p, so the exclusion principle has no application and no conclusion follows to the effect that the initial supposition "m causes p" is false. The real lesson of the argument, therefore, is this: *Either accept serious physicalism, like the psychoneural identity theory, or face the specter of epiphenomenalism!*

As noted, the epiphenomenalism involved here concerns only the efficacy of mental events in the causation of physical events, not the causal power of mentality in general. However, a more radical epiphenomenalism rears its unwelcome head in the next section.

The "Supervenience Argument" and Epiphenomenalism

When you throw mind-body supervenience into the mix, an even more serious threat of epiphenomenalism arises. (The argument can be run in terms of

the idea that mental properties are "realized" by physical-neural properties rather than the premise that the former supervene on the latter.) Let us understand mind-body supervenience in the following form:

> *Mind-Body Supervenience*. When a mental property, M, is instantiated by something *x* at *t*, that is in virtue of the fact that *x* instantiates, at *t*, a physical property, P, such that anything that has P at any time necessarily has M at the same time.

So whenever you experience a headache, that is in virtue of the fact that you are in some neural state N at the time, where N is a supervenience base of headaches in the sense that anyone who is in N must be having a headache. There are no free-floating mental states; every mental state is anchored in a physical-neural base on which it supervenes.

Given mind-body supervenience, an argument can be developed that appears to have disastrous epiphenomenalist consequences.

> (1) Suppose that a mental event, an instantiation of mental property M, causes another mental property, M*, to instantiate.
>
> (2) According to mind-body supervenience, M* is instantiated on this occasion in virtue of the fact that a physical property—one of its supervenience bases—is instantiated on this occasion. Call this physical base P*.
>
> (3) Now ask: Why is M* instantiated on this occasion? What is responsible for the fact that M* occurs on this occasion? There appear to be two presumptive answers: (i) because an instance of M caused M* to instantiate (our original supposition), and (ii) because a supervenience base, P*, of M*, was instantiated on this occasion.

Now, there appears to be strong reason to think that (ii) trumps (i): If its supervenience base P* occurs, M* must occur, no matter what preceded M*'s occurrence—that is, *as long as P* is there, M* is guaranteed to be there even if its supposed cause M did not occur*. This undermines M's claim to have brought about this instance of M*; it seems that P* must take the primary credit for bringing about M* on this occasion. Is there a way of reconciling M's claim to have caused M* to instantiate and P*'s claim to be M*'s supervenience base on this occasion?

(1) The two claims (i) and (ii) can be reconciled if we are willing to accept: M caused M* to instantiate *by causing M*'s supervenience base P* to instantiate*. This seems like the only way to harmonize the two claims.

In general, it seems like a plausible principle to say that in order to cause, or causally affect, a supervenient property, you must cause, or tinker with, its supervenience base. If you are not happy with a painting you have just finished and want to improve it, there is no way you could alter the aesthetic qualities of the painting (for example, make it more expressive, more dramatic, and less sentimental) except by altering the physical properties on which the aesthetic properties supervene. You must bring out your brushes and oils and do physical work on the canvas. That is the only way. You take aspirin to relieve your headache because you hope that ingesting aspirin will bring about physicochemical changes in the neural state on which your headache supervenes.

(2) Hence, M causes P*. This is an instance of mental-to-physical causation.

If this argument is correct, it shows that, given mind-body supervenience, mental-to-mental causation (an instance of M causing M* to instantiate) leads inevitably to mental-to-physical causation. This argument, which may be called the "supervenience argument," shows that mental-to-mental causation is possible only if mental-to-physical causation is possible.

But see where the two arguments, the exclusion argument and the supervenience argument, lead us. According to the supervenience argument, mental-to-mental causation is possible only if mental-to-physical causation is possible. But the exclusion argument says that mental-to-physical causation is not possible. So it follows that neither mental-to-mental causation nor mental-to-physical causation is possible. This goes beyond the epiphenomenalism of mental-to-physical causation; the two arguments together purport to show that mental events have no causal efficacy at all, no power to cause any event, mental or physical. This is radical epiphenomenalism.

It is important to keep in mind that all this holds on the assumption that we do not choose the option of reductionist physicalism; that is, if we reject the premise "$m \neq p^*$" of the exclusion argument, thereby accepting the psychoneural identity "$m = p$," we can avoid the epiphenomenalist conclusion. So the upshot of these two arguments is this: If you want to avoid radical epiphenomenalism, you must be prepared to embrace reductionist physicalism—that is, you must choose between an extreme form of epiphenomenalism and reductionism.

Neither option is palatable. To most of us, epiphenomenalism seems just false, or even incoherent (recall Fodor's lament). And reductionist physicalism does not seem much better: If we save mental causation by reducing mentality to mere patterns of electrochemical activity in the brain, have we really saved mentality as something special and distinctive? Moreover, what if the mental is not reducible to the physical? Aren't we then stuck with epiphenomenalism whether we like it or not? This is the conundrum of mental causation.

The general moral of our discussion seems to be this: If anything is to have causal powers and enter into causal relations with anything else, it must be part of the physical domain. This conclusion complements, and strengthens, what we learned about the problem of mental causation for Descartes's immaterial minds (chapter 2).

Further Issues: The Extrinsicness of Mental States

Computers compute with 0s and 1s. Suppose you have a computer running a certain program, say, a program that monitors the inventory of a supermarket. Given a string of 0s and 1s as input (a can of Campbell's tomato soup has just been scanned at a checkout station), the computer goes through a series of computations and emits an output (the count of Campbell's tomato soup in stock has been adjusted, and so on). Thus, the input string of 0s and 1s represents a can of Campbell's tomato soup being sold, and the output string of 0s and 1s represents the amount of Campbell's tomato soup still in stock. When the manager checks the computer for a report on the available stock of Campbell's tomato soup, the computer "reports" that the present stock is such and such, and it does so *because* "it has been told" (by the checkout scanners) that twenty-five cans have been sold so far today. And this "because" is naturally understood as signifying a causal relation.

But we know that it makes no difference to the computer what the strings of 0s and 1s *mean* or *represent*. If the input string had meant the direction and speed of wind at the local airport or the identification code of an employee, or even if it had meant nothing at all, the computer would have gone through exactly the same computation and produced the same output string. In this case, the output string too would have meant something else, but what is clear is that the "meanings," or "representational contents," of these 0s and 1s are in the eye of the computer programmer or user, not something that is involved in the computational process. Give the computer the same string of 0s and 1s as input, and it will go through the same computation every time and give you the

same output. The "semantics" of these strings is irrelevant to computation; what matters is their shape—that is, their syntax. The computer is a "syntactic engine"; it is driven by the shapes of symbols, not their meanings.

According to an influential view of psychology known as computationalism (or the computational theory of mind), cognitive mental processes are best viewed as computational processes on mental representations (chapter 5). According to it, constructing a psychological theory is like writing a computer program; such a theory will specify, for each input (say, retinal stimulation), the computational process that a cognizer will undergo to produce an output (say, the visual detection of an edge). But what the considerations of the preceding paragraph seem to show is that, on the computational view of psychology, the meanings, or contents, of internal representations make no difference to psychological processes. Suppose a certain internal representation, *i*, represents the state of affairs S (say, that there are horses in the field); having S as its representational content, or meaning, is the semantics of *i*. But if we suppose, as is often done on the computational model, that internal representations form a language-like system (the "language of thought"), *i* must also have a syntax, or formal grammatical structure. So if our considerations are right, it is the syntax of *i*, not its semantics, that determines the course of the computational process starting with *i*. The fact that *i* means that there are horses in the field rather than, say, that there are lions in the field, is of no causal relevance to what other representations issue from *i*. The computational process that *i* initiates will be wholly determined by *i*'s syntactic shape. But doesn't this mean that the *contents* of our beliefs and desires and of other propositional attitudes have no causal relevance for psychological processes?

The point actually is independent of computationalism and can be seen to arise for any broadly physicalist view of mentality. Assume that beliefs and desires and other intentional states are neural states. Each such state, in addition to being a neural state with biological-physical properties, has a specific content (for example, that water is wet, or that the Obamas have a home in Chicago). That a given state has the content it has is a *relational*, or *extrinsic*, *property* of that state, for the fact that your belief is about water, or about the Obamas, is in part determined by your causal-historical associations with water and the Obamas (see chapter 8). Let us consider what this means and why it is so.

Suppose there is in some remote region of this universe another planet, "Twin Earth," that is exactly like our Earth, except for the following fact: On Twin Earth, there is no water, that is, no H_2O, but an observably indistinguishable chemical substance, XYZ, fills the lakes and oceans there, comes out of the tap in Twin Earth homes, and so on. Each of us has a doppelganger there who

is an exact molecular duplicate of us. (Let us ignore the inconvenient fact that your twin has XYZ molecules in her body where you have H_2O molecules.) On Twin Earth, people speak Twin English, which is just like English, except for the fact that their word "water" refers to XYZ, not water, and when they utter sentences containing the expression "water," they are talking about XYZ, not water. Thus, Twin Earth people have thoughts about XYZ, where we have thoughts about water, and when you believe that water is wet, your doppelganger on Twin Earth has the belief that XYZ is wet, even though you and she are molecule-for-molecule duplicates. And when you think that Obama is from Chicago, your twin thinks that the Twin Earth Obama (he is the forty-fourth president of the Twin Earth United States) is from Twin Earth Chicago. And so on. The differences in Earth and Twin Earth belief contents (and contents of other intentional states) are due not to internal physical or mental differences in the believers but to the differences in the environments in which the believers are embedded (see the discussion of "wide content" in chapter 8). Contents, therefore, are extrinsic, not intrinsic; they depend on your causal history and your relationships to the objects and events in your surroundings. States that have the same intrinsic properties—the same neural-physical properties—may have different contents if they are embedded in different environments. Further, an identical internal state that lacks an appropriate relationship to the external world may have no representational content at all.

But isn't it plausible to suppose that behavior causation is "local" and depends only on the intrinsic neural-physical properties of these states, not their extrinsic relational properties? Isn't it plausible to suppose that someone whose momentary neural-physical state is exactly identical with yours will behave just the way you do—say, raise the right hand—regardless of whether her brain state has the same content as yours? This raises doubts about the causal relevance of contents because the properties of our mental states implicated in behavior causation are plausibly expected to be intrinsic. What causes your behavior, we feel, must be *local—in you*, *here and now*; after all, the behavior it is supposed to cause is here and now. But contents of mental states are relational and extrinsic; they depend on what is out there in the world outside you, or on what occurred in the past and is no longer here. To summarize, contents do not supervene on the intrinsic properties of the states that carry them; on the other hand, we expect behavior causation to be local and depend only on intrinsic properties of the behaving organism. This, then, is yet another problem of mental causation. It challenges us to answer the following question: How can intentional mental states, like beliefs and desires, be efficacious in behavior causation in virtue of their contents?

Various attempts have been made to reconcile the extrinsicness of contents with their causal efficacy, but we do not as yet have a fully satisfactory account. The problem has turned out to be a highly complex one involving many issues in metaphysics, philosophy of language, and philosophy of science.[22]

For Further Reading

Donald Davidson's "Mental Events" is the primary source of anomalous monism. On the problem of mental causation associated with anomalous monism, see Ernest Sosa, "Mind-Body Interaction and Supervenient Causation," and Louise Antony, "Anomalous Monism and the Problem of Explanatory Force." Davidson responds in "Thinking Causes," which appears in *Mental Causation*, edited by John Heil and Alfred Mele. This volume also contains rejoinders to Davidson by Kim, Sosa, and Brian McLaughlin, as well as a number of other papers on mental causation.

For counterfactual-based accounts of mental causation, see Ernest LePore and Barry Loewer, "Mind Matters," and Terence Horgan, "Mental Quausation." On functionalism and mental causation, see Ned Block, "Can the Mind Change the World?" and Brian McLaughlin, "Is Role-Functionalism Committed to Epiphenomenalism?"

Journal of Consciousness Studies, vol. 13, no. 1–2, edited by Michael Pauen, Alexander Staudacher, and Sven Walter, is a special issue on epiphenomenalism and contains many interesting papers on the topic.

For issues related to the causal role of extrinsic mental states, see Fred Dretske, "Minds, Machines, and Money: What Really Explains Behavior," and Tim Crane, "The Causal Efficacy of Content: A Functionalist Theory." Many of the issues in this area are discussed in Lynne Rudder Baker, *Explaining Attitudes*; Dretske, *Explaining Behavior*; Pierre Jacob, *What Minds Can Do*. Stephen Yablo's "Wide Causation" is interesting but difficult and challenging.

On the principle of physical causal closure, see David Papineau's "The Rise of Physicalism" and "The Causal Closure of the Physical and Naturalism." For a different perspective, see E.J. Lowe, "Physical Causal Closure and the Invisibility of Mental Causation" and "Non-Cartesian Substance Dualism and the Problem of Mental Causation."

On the exclusion and supervenience arguments, see Jaegwon Kim, *Mind in a Physical World* and *Physicalism, or Something Near Enough*, chapter 2. Many interesting papers on issues on these and related topics are found in *Physicalism and Mental Causation*, edited by Sven Walter and Heinz-Dieter Heckmann. Recommended also are Stephen Yablo, "Mental Causation"; Karen

Bennett, "Why the Exclusion Problem Seems Intractable, and How, Just Maybe, to Tract It" and "Exclusion Again"; John Gibbons, "Mental Causation Without Downward Causation." For an interesting and wide-ranging discussion of the exclusion principle and related issues, see Christian List and Peter Menzies, "Nonreductive Physicalism and the Limits of the Exclusion Principle."

Some philosophers advocate the "trope" theory as basic ontology, in order to get around some of the difficulties with mental causation. A good example is "The Metaphysics of Mental Causation" by Cynthia Macdonald and Graham Macdonald.

Karen Bennett's "Mental Causation" is a balanced and accessible overview and discussion of mental causation.

Notes

1. Marcel Proust, *Remembrance of Things Past*, vol. 1, pp. 48–51.

2. Some philosophers insert another step between beliefs-desires and actions, by taking beliefs-desires to lead to the formation of *intentions* and *decisions*, which in turn lead to actions. What has been described is the influential causal theory of action, which is widely, but far from universally, accepted. Details concerning action, agency, and action explanation are discussed in a subfield of philosophy called action theory, or the philosophy of action.

3. Whether explanations appealing to emotions presuppose belief-desire explanations is a controversial issue. For discussion, see Michael Smith, "The Possibility of Philosophy of Action."

4. Donald Davidson, "Actions, Reasons, and Causes." For noncausal approaches, see Carl Ginet, *On Action*, and Frederick Stoutland, "Real Reasons."

5. Thomas H. Huxley, "On the Hypothesis That Animals Are Automata, and Its History," *Philosophy of Mind: Classical and Contemporary Readings*, ed. David J. Chalmers, pp. 29–30.

6. Huxley advances his epiphenomenalism in regard to consciousness; it isn't clear what his views are about the causal status of mental states like beliefs and desires. Does the French sergeant perform actions? Does he have beliefs and desires?

7. Samuel Alexander, *Space, Time, and Deity*, vol. 2, p. 8.

8. Jerry A. Fodor, "Making Mind Matter More," in Fodor, *A Theory of Content and Other Essays*, p. 156.

9. More precisely, Davidson's claim is that there are no "strict" laws connecting psychological and physical phenomena. There are some questions about

what the strictness of laws amounts to; for our present purposes, it is sufficient to understand "strict" as "exceptionless." See Davidson's "Mental Events."

10. See Donald Davidson's "Mental Events." For an interpretive reconstruction of Davidson's argument, see Jaegwon Kim, "Psychophysical Laws."

11. This is a form of what is called "the principle of charity"; Davidson also requires that an interpretation of a person's belief system make her beliefs come out largely *true*. See the discussion of interpretation theory in chapter 8.

12. In "Mental Events," Davidson defends the stronger thesis that there are no laws at all about mental phenomena, whether psychophysical or purely psychological; his view is that laws (or "strict laws") can be found only in basic physics (see "Thinking Causes"). A sharp-eyed reader will have noticed that Davidson's argument requires this stronger thesis, since the argument as it stands leaves open the possibility that the two causally connected events, *m* and *p*, instantiate a purely psychological law, from which it would follow that *p* is a mental event. If, as Davidson believes, "strict" laws are found only in physics, his conclusion can be strengthened: Any event (of any kind) that causes, or is caused by, another event (of any kind) is a physical event. For a defense of the thesis that there are no laws at all about psychological phenomena, see Jaegwon Kim, "Why There Are No Laws in the Special Sciences: Three Arguments."

13. Brian McLaughlin calls it "type epiphenomenalism" in his "Type Epiphenomenalism, Type Dualism, and the Causal Priority of the Physical." Several philosophers independently raised these epiphenomenalist difficulties for anomalous monism; Frederick Stoutland was probably the first to do so, in his "Oblique Causation and Reasons for Action."

14. This is not quite complete. The counterfactual analysis of causation only requires that there be a chain of these "counterfactual dependencies" connecting cause and effect. But this and other refinements do not affect the discussion to follow. David Lewis's "Causation" is the first full counterfactual analysis of causation.

15. Of late, the possible-world semantics has been dominant for counterfactuals; the first approach has virtually disappeared from the scene.

16. See Ernest Nagel, *The Structure of Science*, chapter 4.

17. For a detailed development of this approach, see David Lewis, *Counterfactuals*. Lewis's account does not require that there be "the closest" P-world; there could be ties.

18. These worlds are very much underdescribed, of course; we are assuming that the worlds are roughly the same in other respects.

19. Later in his career, Davidson too came to accept nonstrict laws as capable of grounding causal relations; see his "Thinking Causes." But this may very well undermine his argument for anomalous monism.

20. See David Papineau, "The Rise of Physicalism."

21. For more detail, see Jaegwon Kim, *Physicalism, or Something Near Enough*, chapter 2.

22. The inability to reach a satisfactory solution to this problem can add fuel to the eliminativist argument on content-carrying mental states, along the lines urged by Paul Churchland in "Eliminative Materialism and the Propositional Attitudes." If contents are causally inefficacious, how can they play a role in causal-explanatory accounts of human behavior? And if they can have no such role, why should we bother with them, whether in commonsense psychology or the science of human behavior?