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 4

Mind as the Brain

The Psychoneural Identity Theory

Some ancient Greeks thought that the heart was the organ responsible for thoughts and feelings—an idea that has survived, we are told, in the traditional symbolism of the heart as signifying love and romance. But the Greeks got it wrong; we now know, as surely as such things can be known, that the brain is where the action is as far as our mental life is concerned. If you ask people where their minds or thoughts are located, they will point to their heads. Does this mean only that the mind and brain share the same location, or something stronger, namely, that the mind *is* the brain? We consider here a theory that advocates this stronger claim—that the mind is identical with the brain and that for a creature to have mentality is for it to have a brain with appropriate structure and capacities.

Mind-Brain Correlations

But what makes us think that the brain is "the seat of our mental life," as Descartes might have put it? The answer seems clear: There are *pervasive and systematic psychoneural correlations*, that is, *correlations between mental phenomena and neural states of the brain*. This is not something we know a priori; we know it from empirical evidence. We observe that injuries to the brain often have a dramatic impact on mental life, affecting the ability to reason, recall, and perceive, and that they can drastically impair a person's cognitive capacities and even alter her personality traits. Chemical changes in the brain brought on by ingestion of alcohol, antidepressants, and other psychoactive drugs affect our moods, emotions, and cognitive functions. When a brain concussion

knocks us out, our conscious life goes blank. Sophisticated brain imaging techniques allow us to "see" just what is going on in our brains when we are engaged in certain mental activities, like seeing green or feeling agitated. It is safe to say that we now have overwhelming scientific evidence attesting to the centrality of the brain and its activities as determinants of our mental life.

A badly scraped elbow can cause you a searing pain, and a mild food poisoning is often accompanied by stomachaches and queasy feelings. Irradiations of your retinas cause visual sensations, which in turn cause beliefs about objects and events around you. Stimulations of your sensory surfaces lead to sensory and perceptual experiences of various kinds. However, peripheral neural events are only remote causes; we think that they bring about conscious experiences only because they cause appropriate states of the brain. This is how anesthesia works: If the nerve signals coming from sensory peripheries are blocked or the normal functions of the brain are interfered with so that the central neural processes that underlie conscious experience are prevented from occurring, there will be no experience of pain—perhaps no experience of anything. It is plausible that everything that occurs in mental life has a state of the brain (or the central nervous system) as its *proximate* physical basis. It would be difficult to deny that the very existence of our mentality depends on the existence of appropriately functioning neural systems: If all the cells and molecules that make up your brain were scattered in intergalactic space, your whole mental life would vanish at that moment, just as surely as annihilating all the molecules making up your body would mean its end. At least that is the way things seem. We may summarize this in the following thesis:

> *Mind-Brain Correlation Thesis*. For each type M of mental event that occurs to an organism *o*, there exists a brain state of kind B (M's "neural correlate" or "substrate") such that M occurs to *o* at time *t* if and only if B occurs to *o* at *t*.

According to this thesis, then, each type of mental event that can occur to an organism has a neural correlate that is both necessary and sufficient for its occurrence. So for each organism there is a set of mind-brain correlations covering every kind of mental state it is capable of having.

Two points may be noted about these brain-mind correlations:

1. They are "lawlike": The fact that pain is experienced when certain of your neurons (say, C-fibers and Aδ-fibers) are activated is a matter of *lawful regularity*, not accidental, or coincidental, co-occurrences.

2. Even the smallest change in your mental life cannot occur unless there are some specific (perhaps still unknown) changes in your brain state; for example, when your headache goes away, there must be an appropriate change in your neural states.

Another way of putting these points, though this is not strictly equivalent, is to say that mentality *supervenes* on brain states. Remember that this supervenience, if it indeed holds, is something we know from observation and experience, not a priori. Moreover, specific correlations—that is, correlations between specific types of mental states (say, pain) and specific types of brain states (say, the activation of certain neural fibers)—are again matters of scientific research and discovery, and we may assume that many of the details about these correlations are still largely unknown. However, it is knowledge of these specific correlations, rough and incomplete though it may be, that ultimately underlies our confidence in the general thesis of mind-brain correlation and mind-brain supervenience. If Aristotle had been correct (and he *might* have been correct) about the heart being the engine of our mentality, we would have a mind-heart correlation thesis and mind-heart supervenience, instead of the mind-brain correlation thesis and mind-brain supervenience.

Making Sense of Mind-Brain Correlations

When a systematic correlation between two properties or types of events has been observed, we want an explanation, or interpretation, of the correlation: Why do the properties F and G correlate? Why is it that an event of type F occurs just when an event of type G occurs? We do not want to countenance too many "brute," unexplained coincidences in nature. An explanatory demand of this kind becomes even more pressing when we observe systematic patterns of correlation between two large families of properties, like mental and neural properties. Let us first look at some examples of property correlations outside the mind-brain case:

a. Whenever the ambient temperature falls below 20 degrees Fahrenheit and stays there for several days, the local lakes and ponds freeze over. Why? The answer, of course, is that the low temperature *causes* the water in the ponds to freeze. The two events are *causally related*, and that is why the observed correlation occurs.
b. You enter a clock shop and find an astounding scene: Dozens and dozens of clocks of all shapes and sizes are busily ticking away, and

they all show exactly the same time, 2:00. Awhile later, you see all of them showing exactly 2:30, and so on. What explains this marvelous correlation among these clocks? It could not be a coincidence, we think. One possible answer is that the shopkeeper synchronized all the clocks, which are all working properly, before the shop opened in the morning. Here, a *common cause*, the shopkeeper's action in the morning, explains the correlations that are now observed; to put it another way, one clock showing 3:30 and another showing the same time are *collateral effects of a common cause*. There are no direct causal relationships between the clocks that are responsible for the correlations.

c. We can imagine a slightly different explanation of why the clocks are keeping the same time: These clocks actually are not very accurate, and some of them gain or lose time markedly every five minutes or so. But there is a little leprechaun whose job is to run around the shop, unseen by the customers, synchronizing the clocks every minute. That is why every time you look, the clocks show the same time. This again is a *common-cause* explanation of a correlation, but it is different from the story in (b) in the following respect: This explanation involves a continued intervention of a causal agent, whereas in (b) a single cause in the past is sufficient. In neither case, however, is there a direct cause-effect relationship between the correlated events.

d. Why do temperature and pressure covary for gases confined in a rigid container? The temperature and pressure of a gas are both dependent on the motions of the molecules that compose the gas: The temperature is the average kinetic energy of the molecules, and the pressure is the momentum imparted to the walls of the container (per unit area) by the molecules colliding with them. Thus, the rise in temperature and the rise in pressure can be viewed as *two aspects* of one and the same underlying microprocess.

e. Why does lightning occur just when there is an electric discharge between clouds or between clouds and the ground? Because lightning simply *is* an electric discharge involving clouds and the ground. There is here only one phenomenon, not two that are correlated with each other, and what we thought were distinct correlated phenomena turn out to be one and the same event, under two different descriptions. Here an apparent correlation turns out to be an *identity*.

f. Why do the phases of the moon (full, half, quarter, and so on) covary with the tidal actions of the ocean (spring tides, neap tides, and so on)? Because the relative positions of the earth, the moon, and the sun determine both the phases of the moon and the combined strength of the gravitational forces of attraction exerted on the ocean water by the moon and the sun. So the changes in gravitational force are the proximate causes of tidal actions, and the relative positions of the three bodies can be thought of as their distal cause. The phases of the moon are merely collateral effects of the positions of the three bodies involved and serve only as an indication of what the positions are (full moon when the earth is between the sun and the moon on a straight line, and so on), having no causal role whatever on tidal actions.

What about explaining, or interpreting, mind-brain correlations? Which of the models we have surveyed best fits the mind-body case? As we would expect, all of these models have been tried. We begin with some causal approaches to the mind-body relation:

Causal Interactionism. Descartes thought that causal interaction between the mind and the body occurred in the pineal gland (chapter 2). He speculated that "animal spirits"—fluids made up of extremely fine particles flowing around the pineal gland—cause it to move in various ways, and these motions of the gland in turn cause conscious states of the mind. Conversely, the mind could cause the gland to move in various ways, affecting the flow of the surrounding animal spirits. This in turn influenced the flow of these fluids to different parts of the body, ultimately issuing in various physiological changes and bodily movements.[1]

"Preestablished Harmony" Between Mind and Body. Leibniz, like many of his great contemporary Rationalists, thought that no coherent sense could be made of Descartes's idea that an immaterial mind could causally influence, or be influenced by, a material body like the pineal gland, managing to move this not-so-insignificant lump of tissue hither and thither. On his view, the mind and the body are in a "preestablished harmony," rather like the clocks that were synchronized by the shopkeeper in the morning, with God having started off our minds and bodies in a harmonious relationship. Whether this is any less fantastical an idea, at least for us, than Descartes's idea of mind-body interaction is debatable.

Occasionalism. According to Nicolas Malebranche, another major Continental Rationalist, whenever a mental event appears to cause a physical event or a physical event appears to cause a mental event, it is only an illusion. There is no direct causal relation between "finite minds" and bodies; when a mental event, say, your will to raise your arm, occurs, that only serves as an *occasion* for God to intervene and cause your arm to rise. Divine intervention is also responsible for the apparent causation of mental events by physical events: When your finger is cut, that again is an occasion for God to step in and cause you pain. The role of God, then, is rather like that of the leprechaun in the clock shop whose job is to keep the clocks synchronized at all times by continuous interventions. This view is known as occasionalism; it was an outcome of the doctrine, accepted by Malebranche and many others at the time, that God is the only genuine causal agent in this world, and that the apparent causal relations we observe in the created world are only that, an appearance.

The Double-Aspect Theory. Spinoza, another great Rationalist of the time, maintained that mind and body are simply two correlated aspects of a single underlying substance that is in itself neither mental nor material. This theory, like the doctrine of preestablished harmony and occasionalism, denies direct causal relationships between the mental and the physical; however, unlike them, it does not invoke God's causal action to explain the mental-physical correlations. The observed correlations are there because they are two distinguishable aspects of one underlying reality. A modern form of this approach is known as neutral monism, according to which the fundamental reality is neutral in the sense that it is intrinsically neither physical nor mental.

Epiphenomenalism. According to T. H. Huxley, a noted British biologist of the nineteenth century, all conscious events are caused by neural events in the brain, but they have no causal power of their own, being the ultimate end points of causal chains.[2] So all mental events are effects of the physiological processes in the brain, but they are powerless to cause anything else—even other mental events. You "will" your arm to rise, and it rises. But to think that your volition is the cause of the rising of the arm is to commit the same error as thinking that the changes in the phases of the moon cause the changes in tidal motions. The real cause of the arm's rising is a certain neural event in your brain, and this event also causes your experience of a volition to raise the arm. This is like the case of the moon and the tides: The relative positions of the earth, the moon, and the sun are the true cause of both the tidal motions

and the phases of the moon. Many scientists in brain research seem to hold, at least implicitly, a view of this kind (see chapter 10).

Emergentism. There is another interesting response to the question "Why are mental phenomena correlated with neural phenomena in the way they are?" It is this: The question is unanswerable—the correlations are "brute facts" that we must simply accept; they are not subject to further explanation. This is the position of emergentism. It holds that when biological processes attain a certain level of organizational complexity, a wholly new type of phenomenon, namely, consciousness and rationality, "emerges," and why and how these phenomena emerge is not explainable in terms of the lower-level physical-biological facts. There is no explanation of why, say, pains rather than itches emerge from C-fiber activations or why pains emerge from C-fiber activations rather than another type of neural state. That there are just these emergence relationships and not others must be accepted, in the words of Samuel Alexander, a leading theoretician of the emergence school, "with natural piety."[3] The phenomenon of emergence must be recognized as a fundamental fact about the natural world. One important difference between emergentism and epiphenomenalism is that the former, but not the latter, acknowledges causal power and efficacy of emergent mental phenomena.

The Psychoneural (or Psychophysical, Mind-Body) Identity Theory. This position, explicitly advanced as a solution to the mind-body problem in the late 1950s, advocates the *identification* of mental states with the physical processes in the brain. Just as there are no bolts of lightning *over and above* atmospheric electrical discharges, there are no mental events *over and above*, or *in addition to*, the neural processes in the brain. "Lightning" and "electrical discharge" are not dictionary synonyms, and the Greeks probably knew something about lightning but nothing about electric discharges; nonetheless, bolts of lightning are just electric discharges, and the two expressions "lightning" and "atmospheric electric discharge" refer to the same phenomenon. In the same way, the terms "pain" and "C-fiber activation" do not have the same dictionary meaning; Socrates knew a lot about pains but nothing about C-fiber stimulation. And yet pains turn out to be the activations of C-fibers, just as bolts of lightning turned out to be electrical discharges. In many ways, mind-brain identity seems like a natural position to take; it is not just that we point to our heads when we are asked where our minds are. Unless you are prepared to embrace Cartesian immaterial mental substances outside physical space, what

could your mind be if not your brain? And what could mental states be if not states of the brain?

* * *

But what are the arguments that support the identification of mental events with brain events? Even if your mind is in your head, your mind and your brain might only share the same space while remaining distinct. So are there good reasons for thinking that the mind *is* the brain? There are three principal arguments for the mind-brain identity theory. These are the simplicity argument, the explanatory argument, and the causal argument. We will see how these arguments can be formulated and defended, and try to assess their cogency. We will then turn to some arguments designed to refute, or at least discredit, the mind-brain identity theory.

The Argument from Simplicity

J. J. C. Smart, whose 1959 essay "Sensations and Brain Processes" had a critical role in establishing the psychoneural identity theory as a major position on the mind-body problem, emphasized the importance of *simplicity* as a ground for accepting the theory.[4] He writes:

> Why do I wish [to identify sensations with brain processes]? Mainly because of Occam's razor. . . . There does seem to be, so far as science is concerned, nothing in the world but increasingly complex arrangements of physical constituents. All except for one place: in consciousness. That is, for a full description of what is going on in a man you would have to mention not only the physical processes in his tissues, glands, nervous system, and so forth, but also his states of consciousness: his visual, auditory, and tactual sensations, his aches and pains. That these should be *correlated* with brain processes does not help, for to say that they are *correlated* is to say that they are something "over and above." . . . So sensations, states of consciousness, do seem to be the one sort of thing left outside the physicalist picture, and for various reasons I just cannot believe that this can be so. That everything be explicable in terms of physics . . . except the occurrence of sensations seems to me frankly unbelievable.[5]

Occam's (or Ockham's) razor, named after the fourteenth-century philosopher William of Ockham, is a principle that urges simplicity as an important

virtue of theories and hypotheses. The following two formulations are among the standard ways of stating this principle:[6]

I. Entities must not be multiplied beyond necessity.
II. What can be done with fewer assumptions should not be done with more.

Principle (I) urges us to adopt the simplest ontology possible, one that posits no unnecessary entities—that is, entities that have no work to do. In mathematics, we deal with natural numbers, rationals, and reals. But real numbers can be constructed out of rationals, which in turn can be constructed out of natural numbers. Natural numbers, too, can be generated as a series of sets. Sets are all we need to do mathematics. A crucial question in applying this principle, of course, is to determine what counts as going "beyond necessity," or what "work" needs to be done. The physicalist would hold that Cartesian immaterial minds are useless and unneeded posits; the Cartesian dualist, however, would disagree precisely on that point.

Principle (II) can be taken as urging simplicity and economy in theory construction: Choose the theory that gives the simplest, most parsimonious descriptions and explanations of the phenomena in its domain—that is, the theory that does its work with the fewest independent hypotheses and assumptions. When Napoleon asked the astronomer and mathematician Pierre de Laplace why God was absent from his theory of the planetary system, Laplace is reported to have replied, "Sir, I have no need of that hypothesis." To explain what needs to be explained (the stability of the planetary system, in this instance), we do well enough with physical laws alone; we need no help, and get none, from the "hypothesis" that God exists. Here, he is invoking version (II) of Ockham's razor. We can also see Laplace as invoking version (I): We don't need God in our ontology to do planetary astronomy; he would be an idler with no work to do.

There seem to be three lines of consideration one might pursue in attempting to argue in favor of the mind-brain identity theory on the ground of simplicity.

First, it is a simple fact that identification reduces the number of putative entities and thereby enhances ontological simplicity. When you say *X* is the same thing as *Y*—or, as Smart puts it, that *X* is nothing "over and above" *Y*—you are saying that there is just one thing here, not two. So if pain as a mental kind is identified with its neural correlate, we simplify our ontology on two levels: First, there is no mental kind, being in pain, in addition to C-fiber

stimulation; second—and this follows from the previous point—there are no individual pain occurrences in addition to occurrences of C-fiber stimulation. In this rather obvious way, mind-brain identification simplifies our ontology.

Second, it may also be argued that psychoneural identification is conducive to conceptual or linguistic simplicity as well. If all mental states are systematically identified with their neural correlates, there is a sense in which mentalistic language—language in which we speak of sensations, emotions, and thoughts—is *in principle* replaceable by a physical language in which we speak of neural processes. The mentalistic language is practically indispensable and we can be certain that it will remain so. We will almost certainly never have a full catalog of mental-neural correlations, and who among us will want to learn the bewilderingly complex and arcane medical terms? Still, we cannot deny the following crucial fact: On the identity theory, descriptions formulated in a mental vocabulary do not report facts or phenomena distinct from those reportable by sentences in a comprehensive physical-biological language. There are no excess facts beyond physical facts that can only be described in some nonphysical language. In this sense, physical language would be complete and universal.

Third, and this is what Smart seems to have in mind, suppose we stop short of identifying pain with C-fiber stimulation and stick with the correlation "Pain occurs if and only if (iff) Cfs occurs." As earlier noted, correlations cry out for explanation. How might such correlations be explained? In science, we standardly explain laws and correlations by deriving them from other, more fundamental laws and correlations. From what more basic correlations could we derive "Pain occurs iff Cfs occurs"? It seems quite certain that it cannot be derived from purely physical-biological laws alone. The simple reason is that these laws do not even speak of pain; the term, or concept, "pain" does not appear in physical-biological laws, for the obvious reason that it is not part of the physical-biological language. So if the pain-Cfs correlation is to be explained, its explanatory premises (premises from which it is to be derived) will have to include at least one law correlating some mental phenomenon with a physical-biological phenomenon—that is, at least one psychoneural correlation. But this puts us back in square one: How do we explain this perhaps more fundamental mental-physical correlation?

The upshot is that we are likely to be stuck with the pain-Cfs correlation and countless other such psychoneural correlations, one for each distinct type of mental state. (Think about how many mental states there are or could be, and in particular, consider this: For each declarative sentence *p*, such as "It will

snow tomorrow," there is the belief that *p*—that is, the belief that it will snow tomorrow.) And all such correlations would have to be taken as "brute" basic laws of the world—"brute" in the sense that they are not further explainable and must be taken to be among the fundamental laws of our total theory of the world. (We will shortly discuss an argument, "explanatory argument I," that claims that these psychoneural correlations are explained by psychoneural identities; for example, that "pain occurs iff Cfs occurs" is explained by "pain = Cfs.")

But such a theory of the world should strike us as intolerably complex and bloated—the very antithesis of simplicity and elegance we strive for in science. For one thing, it includes a huge and motley crowd of psychoneural correlation laws—a potentially infinite number of them—among its basic laws. For another, each of these psychoneural laws is highly complex: Pain may be a "simple" sensory quality, but look at the physical side of the pain-Cfs correlation. Cfs consists of an untold number of molecules, atoms, and particles, and their interactions. We expect our basic laws to be reasonably simple, and reasonably few in number. And we expect to explain complex phenomena by combining and iteratively applying a few simple laws. We do not expect basic laws to deal in physical structures consisting of zillions of particles in unimaginably complex configurations. This makes our total theory messy, inflated, and inelegant.

Compare this bloated picture with what we get if we move from psychoneural correlations to psychoneural identities—from "pain occurs iff Cfs occurs" to "pain = Cfs." Pain and Cfs are one and not two, and we are not faced by two distinct phenomena whose correlation needs to be explained. In this way, psychoneural identities permit us to *transcend* and *renounce* these would-be correlation laws—what Herbert Feigl aptly called "nomological danglers."[7] Moreover, as Smart emphasizes, the identification of the mental with the physical brings the mental within the purview of physical theory, and ultimately our basic physics constitutes a complete and comprehensive explanatory framework adequate for all aspects of the natural world. The resulting picture is far simpler and more elegant than the earlier picture in which any complete theory of the world must include all those complex mind-brain laws in addition to the basic laws of physics. Anyway, that is the argument.

What should we think of this argument? Does going from psychoneural correlations to psychoneural identities really simplify our total theory of the world, as the argument claims? Here the reader is invited to reflect on the following simple question: Doesn't the psychoneural identity theory *merely*

replace psychoneural correlations with an *equal* number of psychoneural identities, one for one? The identities are empirical just like the correlations, and they make even stronger modal assertions about the world, going beyond the correlations. This is so because the identity "pain = Cfs" is now generally taken to be a necessary truth (if true), and the correlation "pain occurs iff Cfs occurs," being entailed by a necessary truth, turns out itself to be a necessary truth. Moreover, these identities are not deducible from more basic physical-biological laws any more than the correlations are, and so they must be countenanced as fundamental and ineliminable postulates about how things are in the world. So don't we end up with the same number of empirical assumptions about the world? The fact is that the total empirical content of a theory with psychoneural identities is at least equal to that of a theory with the psychoneural correlations they replace. Doesn't it follow that version (II) of the simplicity principle actually argues *against* psychoneural identities, or declares a tie between the identities and the correlations? So what exactly are the vaunted benefits of simplification promised by the identities?

The reader is also invited to consider how a Cartesian, or a dualist of any stripe, might respond to Smart's simplicity argument, keeping in mind that one person's "simple" theory may well be another person's "incomplete" or "truncated" theory. What counts as "going beyond necessity" can be a matter of dispute—in fact, what is to be included among "the necessities" is usually the very bone of contention between the disputants.

Explanatory Arguments for Psychoneural Identity

According to some philosophers, psychoneural identities can do important and indispensable explanatory work—that is, they help explain certain facts and phenomena that would otherwise remain unexplained, and this provides us with a sufficient warrant for their acceptance. Sometimes an appeal is made to the principle of "inference to the best explanation." This principle is usually taken as an inductive rule of inference, and there is a widespread, if not universal, agreement that it is an important rule used in the sciences to evaluate the merits of theories and hypotheses. The rule can be stated something like this:

> *Principle of Inference to the Best Explanation*. If hypothesis H gives the *best* explanation of phenomena in a given domain when compared with other rival hypotheses $H_1, \ldots, H_n$, we may accept H as true, or at least we should prefer H over $H_1, \ldots, H_n$.[8]

It is then argued that psychoneural identities, like "pain = Cfs," give the best explanations of certain facts, better than the explanations afforded by rival theories. The conclusion would then follow that the mind-body identity theory is the preferred perspective on the mind-body problem.

This argument comes in two versions, which diverge from each other in several significant ways. We consider them in turn.

Explanatory Argument I

The two explanatory arguments differ on the question of what it is that is supposed to be explained by psychoneural identities—that is, on the question of the "explanandum." Explanatory argument I takes the explanandum to be psychoneural correlations, claiming that psychoneural identities give the best explanation of psychoneural correlations. As we will see, explanatory argument II claims that the identities, rather than explaining the correlations, explain certain other facts about mental phenomena that would otherwise go unexplained. Let us see how the first explanatory argument is supposed to work.

First, it is claimed that specific psychoneural identities, like "pain = Cfs" and "consciousness = pyramidal cell activity," explain the corresponding correlations, like "pain occurs iff Cfs occurs" and "a person is conscious iff pyramidal cell activity is going on in the brain." As an analogy, consider this: Someone might be curious why Clark Kent turns up whenever and wherever Superman turns up. What better, or simpler, explanation could there be than the identity "Clark Kent *is* Superman"?[9] So the proponents of this form of explanatory argument claim that the following is an explanation of a psychoneural correlation and that it is the best available explanation of it:

(α) Pain = Cfs.
Therefore, pain occurs iff Cfs occurs.

Similarly for other psychological properties and their correlated neural properties.

Second, it is also claimed that the psychoneural identity theory offers the best explanation of the pervasive fact of psychoneural correlations, like this:

(β) For every mental property M there is a physical property P such that M = P.
Therefore, for every mental property M there is a physical property P such that M occurs iff P occurs.[10]

If we could show that psychoneural identities are the best explanations of psychoneural correlations, the principle of inference to the best explanation would sanction the conclusion that we are justified in taking psychoneural identities to be true, and that the psychoneural identity theory is the preferred position on the mind-body problem. Anyway, that is the idea.

But does the argument work? Obviously, specific explanations like (α) are crucial; if they do not work as explanations, there is no chance that (β), the explanation of the general mind-correlation thesis, will work. So is (α) an explanation? And is it the best possible explanation of the correlation? A detailed discussion of the second question would be a lengthy and time-consuming business: We would have to compare (α) with the explanations offered by epiphenomenalism, the double-aspect theory, the causal theory, and so on. But we can say this much in behalf of (α): It is ontologically the simplest. The reason is that all these other theories are dualist theories, and in consequence they have to countenance more entities—mental events in addition to brain events. But is (α) overall the *best* explanation? Fortunately, we can set aside this question because there are serious reasons to be skeptical about its being an explanation at all. If it is not an explanation, the question of whether it is the best explanation does not arise.

First consider this: If pain indeed is identical with Cfs, in what sense do they "correlate" with each other? For there is here only one thing, whether you call it "pain" or "Cfs," and as Smart says in the paragraph quoted earlier, you cannot correlate something with itself. For Smart, the very point of moving to the identity "pain = Cfs" is to transcend and cancel the correlation "pain occurs iff Cfs occurs." This is the "nomological dangler" to be eliminated. For it seduces us to ask wrongheaded and unanswerable questions like "Why does pain correlate with Cfs?" "Why doesn't itch correlate with Cfs?" "Why does any conscious experience correlate with Cfs?" and so on. By opting for the identity, we show that these questions have no answers, since the *presupposition* of the questions—namely, that pain *correlates* with Cfs—is false. The question "Why is it the case that *p*?" presupposes that *p* is true. When *p* is false, the question has no correct answer and it cancels itself as an explanandum. Showing that a demand for an explanation rests on a false presupposition is one way to deal with it; providing an explanation is not the only way.

A defender of the explanatory argument might protest our talk of "correlations," objecting that we are assuming, with Smart, that a "correlation" requires two distinct items. We should stop calling "pain occurs iff Cfs occurs" a *correlation*, if that is going to lead anyone to infer pain and Cfs to be two things. It is pointless to be hung up on the word "correlation." Whatever you call it, the fact

expressed by "pain occurs iff Cfs occurs" is explained by the identity "pain = Cfs," and, moreover, this is the best possible explanation of it. That is all we need to make the explanatory argument work.

It is doubtful, however, that this reply will get the explanatory argument out of trouble. In the first place, this move will not make questions like "Why does pain, not itch, correlate with Cfs?" go away. For we can readily reformulate it as follows: Why is it the case that pain occurs iff Cfs occurs, rather than itches occurring just when Cfs occurs? Would we take the following answer from the proponent of the explanatory argument as an acceptable explanation? "That's because pain is identical with Cfs but itch isn't identical with it." It is doubtful that most of us would consider this an informative answer—an informative explanation of why pains, but not itches, are associated with Cfs. Some notable thinkers, William James and T.H. Huxley among them, have long despaired of our ever being able to explain why these particular mind-body associations (or whatever you wish to call them) hold. The idea that simply by moving from mere associations to identities, we can resolve the explanatory puzzles of Huxley and James seems too good to be true.

Second, if it is true that pain = Cfs, the fact to be explained, namely that pain occurs iff Cfs occurs, is just the fact that pain occurs iff pain occurs, or that Cfs occurs iff Cfs occurs, and these manifestly trivial facts (if they are facts at all), with no content, seem neither in need of an explanation nor capable of receiving one. So rather than offering an explanation of why pain occurs just in case Cfs occurs, the proposal that pain = Cfs transforms the supposed explanandum into something for which explanation seems entirely irrelevant. Rather than explaining it, it disqualifies it as an explanandum.

As we have seen, the argument under consideration invokes the principle of inference to the best explanation as a scientific rule of induction; however, most explanations of correlations in the sciences seem to work quite differently. There appear to be two common ways of explaining correlations in science. First, scientists sometimes explain a correlation by deducing it from more fundamental correlations and laws (as when the correlation between the length and the period of swing of a simple pendulum is explained in terms of more basic laws of mechanics). Second, a correlation is often explained by showing that the two correlated phenomena are collateral effects of a common cause. (Recall the earlier example in which the correlation between the phases of the moon and tidal actions is explained in terms of the astronomical configurations involving the sun, the moon, and the earth; an explanation of co-occurrences of two medical symptoms on the basis of a single underlying disease.) It should be noticed that neither of these two ways renders the correlations into trivialities;

these explanations respect their status as correlations and provide serious and informative explanations for them. Indeed, it is difficult to think of a scientific example in which a correlation is explained by simply identifying the phenomena involved.

There is a further notable feature of scientific hypothesis testing: When a new hypothesis is proposed as the best explanation of the existing data, the scientists do not stop there; they will go on to subject the hypothesis to further tests, by deriving additional predictions and looking for new applications. When "pain = Cfs" is proposed as the best explanation of "pain occurs iff Cfs occurs," what *further* predictions can we derive from "pain = Cfs" for additional tests? Are there predictions, empirical or otherwise, derivable from this identity that are not derivable from the correlation "Cfs causes pain," or the emergent hypothesis "pain is an emergent phenomenon arising from Cfs," or the epiphenomenalist hypothesis "Cfs causes pain"? It seems clear that genuine scientific uses of the inference to the best explanation principle bears little resemblance to its use in explanatory argument I for psychoneural identities. The principle of inference to the best explanation gains credibility from its use in scientific hypothesis testing. Using it to support what is an essentially philosophical claim, with no predictive implications of its own and hence no possibility of further tests, seems at best a misapplication of the principle; it can mislead us into thinking that the choice of a position on the mind-body problem is like a quotidian testing of rival scientific hypotheses. Even J. J. C. Smart, arguably the most optimistic and stalwart physicalist ever, had this to say:

> If the issue is between (say) a brain-process thesis and a heart thesis, or a liver thesis, or a kidney thesis, then the issue is a purely empirical one, the verdict is overwhelmingly in favor of the brain. . . . On the other hand, if the issue is between a brain-or-liver-or-kidney thesis (that is some form of materialism) on the one hand and epiphenomenalism on the other hand, then the issue is not an empirical one. For there is no conceivable experiment which could decide between materialism and epiphenomenalism.[11]

Further, the following consideration will reinforce our claim that the arguments against explanatory argument I has nothing to do with exploiting an informal connotation of the word "correlation." Let us ask: Exactly how does (α) work as an explanation? Explanation is most usefully thought of as derivation—a logical derivation, or proof, of the explanandum from the explanatory premises. So, then, how might the conclusion "pain occurs iff Cfs occurs" be derived from "pain = Cfs"? In formal logic, there is no rule of inference that says "From

'X = Y' infer 'X occurs iff Y occurs'"—for good reason, since a nonlogical term like "occur" is not part of formal logic. Instead, what we standardly find are the following two rules governing identity:

Axiom schema: X = X
Substitution rule: From " . . . X . . . " and "X = Y," infer " . . . Y . . . "

The first rule says that in a proof you can always write down as an axiom any sentence of the form "X = X," like "Socrates = Socrates" and "3 + 5 = 3 + 5." The second rule allows you to put "equals for equals." To put it another way, if X = Y and something is true of X, the same thing must be true of Y. This is the rule that is of the essence of identity. These two rules suffice to fix the logical properties of identity completely.

The following seems to be the simplest, and most natural, way of deriving "pain occurs iff Cfs occurs" from "pain = Cfs":

(γ) Pain = Cfs.
Pain occurs iff pain occurs.
Therefore, pain occurs iff Cfs occurs.

The first line is the premise, a psychoneural identity. The second line is a simple tautology of sentential logic, an instance of "*p* iff *p*," where *p* is any sentence you please, and we may write down a tautology anywhere in a derivation. The third line, the desired correlation, is derived by substituting "Cfs" for the second occurrence of "pain" in this tautology, in accordance with the substitution rule. As you see, the work that the identity "pain = Cfs" does is to enable us to *rewrite* the contentless tautology, "pain occurs iff pain occurs," by putting equals for equals. That is, the conclusion "pain occurs iff Cfs occurs," is a mere rewrite of "pain occurs iff pain occurs" and is equally contentless. As a mere rewrite rule in (γ), the identity "pain = Cfs" does no explanatory work, and hence cannot earn its warrant from the rule of inference to the best explanation.

If you think that calling the identity a "rewrite rule" is off the mark, trivializing its explanatory contributions, never mind what work the identity does in (γ); just consider this question: Does this derivation look to you like an explanation, a real explanation of anything? Now that you have (γ) in hand, would you say to yourself, "Now I finally understand why pain, not itch, occurs just in case my C-fibers are stimulated. I should tell my neuroscience professor about my discovery tomorrow!"? It seems as though once you recognize the pain-Cfs correlation as something to be explained, something you want to understand,

saying that they are one and the same thing will not meet your explanatory need. You will still wonder why pain, not itch, is identical with Cfs—which seems to take you back to the original question: Why does pain, not itch, co-occur with Cfs?

The role of identities in explanations is not well understood; there has been little informative discussion of this issue in the literature. Further, the view that explanation is fundamentally, or always, a derivational process is not universally accepted. However, the concept of explanation is deeply complex and difficult to pin down, and viewing explanatory processes as consisting in derivational activities is one of the few reasonably firm handles we have on this concept. If the defender of the explanatory argument insists that the explanation she has in mind of "pain occurs iff Cfs occurs" in terms of "pain = Cfs" does not proceed as a derivation, she is welcome to tell us exactly how she conceives of her explanation. That is, she needs to tell us just how the identity manages to explain its associated correlation.

There are reasons, then, to remain unpersuaded by the claim that psychoneural identities explain psychoneural correlations, and that for this reason the identities should be accepted as true.

Explanatory Argument II

This version of the explanatory argument does not claim that mind-body identities explain mind-body correlations; rather, they enable us to explain certain facts about mentality that would otherwise remain unexplained. How might we explain the fact that pain causes a feeling of distress? What is the causal mechanism involved? Suppose we have available the following psychoneural identities:

Pain = Cfs.
Distress = neural state N.

We might then be able to formulate the following neurophysiological explanation of why pain causes distress:

(θ) Neurophysiological laws
Cfs causes neural state N.
(I_1) Pain = Cfs.
(I_2) Distress = neural state N.
Therefore, pain causes distress.

Neurophysiological laws explain why Cfs causes N, and from this we derive our explanandum "Pain causes distress," by putting equals for equals on the basis of the psychoneural identities, (I_1) and (I_2). These identities help us explain a psychological regularity in terms of its underlying neural mechanism, and this seems just the kind of deeper scientific understanding we seek about higher-level psychological regularities.

Compare this with the situation in which we refuse to enhance correlations into identities. The best we could do with correlations would be something like this:

(λ) Neurophysiological laws
Cfs causes neural state N.
(C_1) Pain occurs iff Cfs occurs.
(C_2) Distress occurs iff neural state N occurs.
Therefore, pain correlates with a phenomenon that causes a phenomenon with which distress correlates.

This is no explanation of why pain causes distress; it doesn't even come close. To explain it, we need identities (I_1) and (I_2); correlations (C_1) and (C_2) will not do. According to the friends of this form of the explanatory arguments, an explanatory role of the kind played by psychoneural identities, as in (θ), yields sufficient justification for their acceptance.

Ned Block and Robert Stalnaker, proponents of the explanatory argument of this form, agree with J. J. C. Smart in regarding identities not as explaining their associated correlations but as helping us to get rid of them. They put the point this way:

> If we believe that heat is correlated with but not identical to molecular kinetic energy, we should regard as legitimate the question why the correlation exists and what its mechanism is. But once we realize that heat is molecular kinetic energy, questions like this will be seen as wrongheaded.[12]

Similarly, for "pain occurs iff Cfs occurs" and "pain = Cfs." The identity helps us avoid the "wrongheaded" question "Why does pain correlate with Cfs, not with something else?" by ridding us of the correlation. It is clear that contrary to the claims of explanatory argument I, Block and Stalnaker do not believe that this improper question is answered by the identity "pain = Cfs." We may summarize Block and Stalnaker's argument in favor of psychoneural identities as

follows: These identities *enable* desirable psychological explanations while *disabling* the improper demands for explanation of psychoneural correlations.[13]

How good is this argument? Unfortunately, not very good: The argument turns out to be problematic, for reasons similar to those that made explanatory argument I questionable. The trouble is that in both arguments the identities in question do not seem to do any explanatory work and hence are not qualified to benefit from the principle of inference to the best explanation. We can accept the claim that derivation (θ) gives a neurophysiological explanation of why pain causes distress: Laws of neurophysiology directly explain why Cfs causes neural state N, and given the identities "pain = Cfs" and "distress = neural state N," we would be justified in claiming that neurophysiological laws explain the fact that pain causes distress. This is so because, given the two identities, the statements "pain causes distress" and "Cfs causes neural state N" state one and the same fact. There is here one fact described in two ways—in the vernacular vocabulary and in the scientific vocabulary.

This shows just what goes wrong with explanatory argument II: The identities "pain = Cfs" and "distress = neural state N" do *no explanatory* work in this derivation. Their role is to enable us to *redescribe* a fact that has already been explained. The explanatory activity is over and finished at the second line when "Cfs causes neural state N" has been derived from, and thereby explained by, laws of neurophysiology. What the identities do is allow us to *rewrite* "Cfs causes neural state N" as "pain causes distress," by putting equals for equals. This is useful in presenting our explanatory accomplishment in neuroscience in the familiar "folk" language, but this involves no *explanatory* activity. The verdict, therefore, seems inescapable: Since the psychoneural identities have no involvement in explanation, they are ineligible as beneficiaries of the principle of inference to the best explanation. If there is a beneficiary of this principle in this situation, it is the laws of neuroscience because they do the explanatory work!

Our conclusion, therefore, has to be that both forms of the explanatory argument are vulnerable to serious objections. Their shared weakness is a lack of clear appreciation of just what role the psychoneural identities play in the explanations in which they supposedly figure. Our main contention has been that both arguments invoke, but misapply, the rule of inference to the best explanation, a principle that itself is far from uncontroversial.

An Argument from Mental Causation

By mental causation we mean any causal relation involving a mental event. A pin is run into your palm, causing you a sharp pain. The sudden pain causes

you to cry out and quickly pull back your hand. It also causes a feeling of distress and a desire to be rid of it. Causal relations involving mental and physical events are familiar facts of our everyday experience.

But pains do not occur without a physical basis; let us assume that pains are lawfully correlated with neural state N. So the sharp pain that caused the withdrawal of your hand has an occurrence of N as its neural substrate. Is there any reason for not regarding the latter, a neural event, as a cause of your hand's jerky motion?

Suppose we try to trace the causal chain backward from your hand's movement. The jerky motion was presumably caused by the contraction of muscles in your arm, which in turn was caused by neural signals reaching the muscles. The movement of neural signals is a complex physical process involving electrochemical interactions, and if we keep tracing the series of events backward to its source, we can expect it to culminate in a region in the central nervous system, perhaps in the cortex. Now ask yourself: Will this chain ever reach, or go through, a mental experience of pain, the pain you experienced when the pin was stuck in your palm? What could the transition from a neural event to a nonphysical, private pain event be like? Or the transition from a private pain experience to a public physicochemical neural event? How can a pain experience affect the motion of even a single molecule—speeding it up or slowing it down, or changing its direction? How can that happen? Is it even conceivable? It boggles our imagination!

The chances are that the causal chain culminating in your hand's jerky movement, when traced backward, will completely bypass your pain; there will be more and more neural-physical events as you keep going back, but no mental experiences. Nor does it make sense to postulate a purely mental causal chain, independent of the neural-physical chain, somehow reaching your muscles. (That's known as telekinesis—an alleged "psychic" phenomenon involving a mind causing a physical change at a distance, like bending a spoon by intensely gazing at it.) It seems, then, that the only way to salvage the pain as a cause of your hand motion is to think of it as a neural event. Which neural event? The best and most natural choice is its neural substrate, N (as we supposed), the state that is necessary and sufficient for the occurrence of the pain. This in brief is the causal argument, somewhat informally presented, for identifying mental states, especially states of consciousness, with neural states.

There is a more systematic, and currently influential, version of the causal argument that will now be presented. It begins with a premise asserting that mental causation is real:

i. Mental phenomena have effects in the physical world.

In this context, we take (i) as uncontroversial. Our beliefs and desires surely have the power to move our limbs and thereby enable us to cause things around us to be rearranged—moving the books from my desk to the bookshelves, emptying a waste basket, digging my car out of a snowbank, and starting an avalanche. If our mental states had no causal powers to affect physical things and events around us, we would cease to be agents, only helpless spectators of the passing scene. If that were true, our self-conception of ourselves as effective agents in the world would suffer a complete collapse.

Here is the second premise:

ii. [The causal closure of the physical domain] The physical world is causally closed. That is, if any physical event is caused, it has a sufficient physical cause (and a wholly physical causal explanation).

According to this principle, the physical world is causally self-contained and self-sufficient. It doesn't say that every physical cause has a sufficient physical cause—that is the principle of physical causal determinism. So (ii) is compatible with indeterminism about physical events. What (ii) says is that for any physical event, if we were to trace its causal ancestry, this need never take us outside the physical world. If a physical event has no physical cause, then it has no cause at all and no causal explanation. Further, this principle is compatible with dualism and other forms of nonphysicalism: As far as it goes, there could be a Cartesian world of immaterial minds, alongside the physical world, and all sorts of causal relations could hold in that world. The only thing, according to physical causal closure, is that the physical world must be causally insulated from such worlds; there can be no injection of causal influence into the physical world from outside. This means that there can be no "miracles" brought about by some transcendental, supernatural causal agents from outside physical space-time.

On Descartes's interactionist dualism, the physical causal closure fails: When an immaterial soul makes the pineal gland vibrate, thereby setting in motion a chain of bodily events, the motion of the pineal gland is caused, but it has no physical cause and no physical explanation. And this means that our physical theory would remain forever incomplete in the sense that there are physical events whose occurrences cannot be physically explained. A complete theory of the physical world would require references to nonphysical, immaterial causal agents and forces.

Why should we accept the causal closure of the physical domain? We will enumerate some reasons here without going into great detail.[14] First, there is

the widely noted success of modern science, in particular theoretical physics, which we take to be our basic science. Physics is all-encompassing: Nothing in the space-time world falls outside its domain. If a physicist encounters a physical event for which there is no ready physical explanation, or physical cause, she would consider that as indicating a need for further research; perhaps there are as-yet undiscovered physical forces. At no point would she consider the possibility that some nonphysical force outside the space-time world was the cause of this unexplained physical occurrence. The same seems to be true of research in other areas of science—broadly physical science including chemistry, biology, geology, and the like. If a brain scientist finds a neural event that is not explainable by currently known facts in neural science, what is the chance that she would say to herself, "Maybe this is a case of a Cartesian immaterial mind interfering with neural processes, messing up my experiment. I should look into that possibility!" We can be sure that would never happen. What would such research, investigating the workings of immaterial souls, look like? Where would you start? It isn't just that the principle of physical causal closure is the operative assumption in scientific research—remember that in science success is what counts. It may well be that there is a conceptual incoherence in the idea that there are nonphysical causal forces outside space-time that can causally intervene in what goes on in the space-time world.[15]

From these two premises, (i) and (ii), we have the desired conclusion:

(i) Mental phenomena are physical phenomena.

You might point out, rightly, that the only proposition we are entitled to derive is that only those mental phenomena that cause physical events are physical events.[16] Strictly speaking, that is correct, but remember this: Causation is transitive—that is, if one event causes another, and this second event causes a third, then the first event causes the third. If a mental event causes another mental event, which causes a physical event, the first event causes this physical event, and our argument pronounces it to be a physical event. Such chains of mental events can be as long as you wish; as long as a single event in this chain causes a physical event, every event preceding it in the chain qualifies as a physical event. This should pretty much cover all mental events; it is hard to imagine a mental causal chain consisting exclusively of mental events not touching anything physical anywhere. Even if there were such exceptions, the main physicalist point is made. A qualified conclusion stands: Mental events that have effects in the physical domain are physical events. The pain that causes your hand to pull back in a jerky motion and makes you cry

"Ouch!" is a physical event. But which physical event? What better candidate is there than the brain state that is the neural correlate of pain, namely Cfs? Cfs is a necessary and sufficient condition for the occurrence of pain, and it occurs exactly at the same time as the pain.

If in spite of these considerations you still want to insist on the pain as a separate cause of the hand movement, think of a new predicament in which you will find yourself. For the hand movement would now appear to have two distinct causes, the pain and its neural correlate Cfs, each presumably sufficient to bring it about. Doesn't that make this (and every other case of mental-to-physical causation) a case of causal overdetermination, an instance in which two independent causes bring about a single effect? Given that the hand withdrawal has a sufficient physical cause, namely Cfs, what *further* causal contribution can the pain make? There seems no leftover causal work that the pain has to be called on to perform. Again, the identification of the pain with Cfs appears to dissolve all these puzzles. There is, of course, the epiphenomenalist solution: Both the hand withdrawal and the pain are caused by Cfs, and the pain itself has no further causal role in this situation. But unlike the identity solution, the epiphenomenalist move renders the pain causally inert and ends up rejecting our initial assumption that a sharp pain caused the hand's jerky motion.

Perhaps a reconsideration of that assumption may be in order. The identification of a conscious pain experience with some molecular physical processes in the brain strikes some people as totally incredible and still others as verging on incoherence. If given a choice between taking pain and other experiences as physical processes in the brain on one hand and their causal impotence on the other, some may well consider the latter a preferable option. At this point, what the causal argument does is to give us a choice between psychoneural identity and epiphenomenalism: If you want to protect mental events from epiphenomenalism, you had better identify them with physical processes in the brain. To some, this may seem tantamount to discarding what is distinctively mental in favor of molecular physical processes in the body. On the other hand, if you are unwilling to embrace psychophysical identity, you put the causal powers of mentality in jeopardy. What good is our mentality if it is epiphenomenal? We will return to some of these issues later (chapter 7).

Against Psychoneural Identity Theory

There are three main arguments against the mind-brain identity theory. They are the epistemological argument, the modal argument, and the multiple realization argument. We consider each in turn.

The Epistemological Argument

Epistemological Objection 1. There is a group of objections based on the thought that the mental and the physical differ in their epistemological properties. Let us begin with the simplest, and rather simplistic, one. Medieval peasants knew lots about pains but nothing about C-fibers, and in fact little about the brain. So how can pains be identical with C-fiber excitations?

This objection assumes that the two statements "S knows something about X" and "X = Y" together entail "S knows something about Y." But is this true? It appears false: The same peasants knew a lot about water but nothing about H_2O. But that doesn't make the identity "water = H_2O" false. Suppose the objector persists: The peasants did know something about H_2O; after all, they knew a lot about water, and water *is* H_2O! How should we respond? Perhaps there is a sense in which the medieval peasants knew something about H_2O—we can concede that—but this must be a sense of knowing in which it is possible to know something about X without having the concept of X, or the ability to use the concept in forming thoughts or making judgments, or to use the expression "X" to express beliefs. But in this pale sense of knowing, there would be nothing wrong about saying that the peasants knew something about C-fiber excitation. They knew about C-fiber excitation in the same harmless sense in which they knew about H_2O. So the objection fails.

Epistemological Objection 2. According to the identity theory, specific psychoneural identities (for example, "pains are C-fiber excitations") are empirical truths discovered through scientific observation and theoretical research. If "$D_1 = D_2$" is an empirical truth, the two names or descriptions, D_1 and D_2, must have *independent criteria of application*. Otherwise, the identity would be a priori knowable; consider, for example, identities like "bachelor = unmarried adult male" and "the husband of Xanthippe = Xanthippe's male spouse." When an experience is picked out by a subject as a pain rather than an itch or tingle, the subject must do so by *recognizing*, or *noticing*, a certain distinctive felt character, a "phenomenal" or experiential quality, of the occurrence—its painful, hurtful quality. If pains were picked out by neurophysiological criteria (say, if we used C-fiber excitation as the criterion of pain), the identity of pain with a neural state could not be empirical; it would simply follow from the very criterion governing the concept of pain. This means, the objection goes, that to make sense of the supposed *empirical* character of psychoneural identities, we must acknowledge the existence of phenomenal, qualitative characters of experience distinct from neural properties.[17]

It seems, therefore, that the psychoneural identity physicalist still has these qualitative, phenomenal features of experience to contend with; even to make sense of her theory, there must be these nonphysical, qualitative properties by which we identify conscious experiences. It seems that she must somehow show that subjects do not identify mental states by noticing their qualitative features. Could the type physicalist argue that although a person does identify her experience by noticing its qualitative phenomenal features, they are not irreducible, since phenomenal properties as mental properties are identical, on her view, with physical-biological properties? But this reply is not likely to satisfy many people; it will invite the following response: "But surely when we notice our pains as pains, we do not do that by noticing biological or neural features of our brain states!" We immediately distinguish pains from itches and tickles; if we identified our experiences by their neurophysiological features, we should be able to tell which neurophysiological features represent pain, which represent itches, and so on. But is this credible?

Some philosophers have tried to respond to this question by analyzing away phenomenal properties. For example, Smart attempts to give phenomenal properties "topic-neutral translation."[18] According to him, when we say, "Adam is experiencing an orangish-yellow afterimage," the content of our report may be conveyed by the following "topic-neutral" translation—topic-neutral because it says nothing about whether what is being reported is mental or physical:

> Something is going on in Adam that is like what goes on when he is looking at an orangish-yellow color patch illuminated in good light.

(We suppose "looking" is explained physically in terms of his being awake, his eyes' being open and focused on the color patch, and so on.) Smart would add that this "something" that is going on in Adam is a brain state.

But will this satisfy someone concerned with the problem of explaining how someone manages to identify the kind of experience she is having? There is perhaps something to be said for these translations if we approach the matter strictly from the third-person point of view. But when you are reporting your own experience by saying, "I have a sharp pain in my left thumb," are you saying something like what Smart says that you are? To know that you are having an orangish-yellow afterimage, do you need to know anything about what generally goes on whenever you look at orangish-yellow color patches?

A more recent strategy that has become popular with latter-day-type physicalists is to press *concepts* into service and have them replace talk of properties

in the foregoing objection. The main idea is to concede *conceptual* differences between the mental and the neural but deny that these differences point to ontological differences, that is, differences in the properties to which these concepts apply or refer. This way of attempting to meet the objection is called the "phenomenal concept strategy." When we say that a person notices a pain by noticing its painfulness, this does not mean that the pain has the *property* of painfulness; rather, it means that she is "conceptualizing" her experience under the phenomenal *concept* of being painful—but the experience so conceptualized remains a neural state. The phenomenal concept is not a neural or physical concept; in particular, it is not identical with the concept of C-fiber stimulation. There is no consensus on what phenomenal concepts are; some take them as a type of "recognitional concept," like the concept red, which we apply to things on the basis of direct acquaintance with them; others take them to be a kind of demonstrative concept, like "*this* kind of experience," demonstratively referring to an experience of pain; there are many other views.[19] The main point is that a single property, presumably a physical-neural property, is picked out by both a phenomenal and a neural concept. Thus, we have a dualism of concepts, mental and physical, but a monism of properties, the entities referred to by these concepts. The advantage of framing the issues in terms of phenomenal concepts rather than phenomenal properties is supposed to derive from the fact that properties, whether phenomenal or of other sorts, are "out there" in the world, whereas concepts are part of our linguistic-conceptual apparatus for representing and describing what is out there. The strategy, then, is to take the phenomenal-neural differences out of the domain of facts of the world and bring them into the linguistic-conceptual domain. This, at any rate, is a move that has been made by some physicalists and it is currently receiving much attention in the field. Whether it is an essentially verbal ploy or something that is more substantial remains to be seen.

Epistemological Objection 3. Your knowledge that you are thinking about an upcoming trip to East Asia is direct and private in the way that only first-person knowledge of one's own mental states can be. Others have to make inferences based on evidence and observation to find out what you are thinking, or even to find out that you are thinking. But your knowledge is not based on evidence or inference; somehow you directly know. In contrast, you have no such privileged access to your brain states. Your neurologist and neurosurgeon have much better knowledge of your brain than you do. In brief, mental states are directly accessible by the subject; brain states—and physical states in general—are not so accessible. So how can mental states be brain states?

We should note that for this objection to work, it is not necessary to claim that the subject has *infallible* access to all her mental states. For one thing, infallibility or absolute certainty is not the issue; rather, the issue is *private direct* access—that is, first-person access not based on inference from evidence or observation, the kind of access that no other person has. For another, it is only necessary that the subject have such access to at least *some* of her mental states. If that is the case, these mental states, according to this argument, cannot be identified with brain states, states for which public access is possible.

The identity theorist has to deny either the claim that we have direct private access to our own current mental states or the claim that we do not have such access to our brain states. She might say that when we know that we are in pain, we do have epistemic access to our Cfs, but our knowledge is under the description, or concept, "pain," not under the description "Cfs." Here there is one thing, Cfs (that is to say, pain), that can be known under two "modes of presentation"—pain and Cfs. Under one mode, the knowledge is private; under the other, it is public. It is like the same person is known both as "the husband of Xanthippe" and "the drinker of hemlock." You may know Socrates under one description but not the other. So knowledge is relative to the mode of description or conceptualization. Certain brain states, like Cfs, can be known in two different modes or under two different sorts of concepts, mental and physical. Knowledge under one mode can be different from knowledge under the other, and they need not co-occur. So this reply is in line with the final physicalist reply to epistemological objection 2, discussed earlier, which invoked phenomenal concepts. These replies, therefore, will likely stand or fall together.

In considering the viability of this reply, we can grant the point that knowledge and belief do depend on "modes of presentation" or ways of conceptualization or description. This seems like a plausible, and true, claim. What we ought to press for answers and elucidation is the following group of questions: Why is there a class of concepts or modes of presentation that gives rise to a very special type of knowledge, that is, knowledge by direct private access? There seems to be a philosophically important difference between such knowledge and our sundry knowledge of physical objects and events. What characteristics of this distinguished class of concepts and these modes of presentation explain the fact that they allow this special type of knowledge? If we conceptualize C-fiber stimulation under the mental concept "itch," that would presumably be wrong. Why? What makes it wrong? The dualist seems to have a simple perspective on these issues: These mental concepts and modes of presentation apply to, or signify, mental events that are directly and privately accessible to the subject; there is not, nor need there be, anything special about the concepts and

modes of presentation themselves. This is exactly the kind of reply that the psychoneural identity theorist wants to avoid.

The Modal Argument

Type physicalists used to say that mind-brain identities—for example, "pain = C-fiber activation"—are *contingent*, not necessary. That is, although pain is in fact C-fiber excitation, it could have been otherwise; there are possible worlds in which pain is not C-fiber excitation but some other brain state—perhaps not a brain state at all. The idea of contingent identity can be explained by an example such as this: "Barack Obama is the forty-fourth president of the United States." The identity is true, but it might have been false: There are possible worlds in which the identity does not hold—for example, one in which Obama decided to pursue an academic career rather than politics, one in which Senator Hillary Clinton won the Democratic nomination, one in which Senator John McCain defeated Obama, and so on. In all these worlds someone other than Barack Obama would be the forty-fourth president of the United States.

But this is possible only because the expression "the forty-fourth president of the United States" can refer to different persons in different possible worlds; things might have gone in such a way that the expression designated someone other than Obama—for example, Hillary Clinton or John McCain. Expressions like "the forty-fourth president of the United States," "the 2009 Wimbledon Men's Singles Champion," and "the tallest man in China," which can name different things in different possible worlds, are what Saul Kripke calls "nonrigid designators."[20] In contrast, proper names like "Barack Obama," "Socrates," and "Number 7" are "rigid"—they designate the same objects in all possible worlds in which they exist. The forty-fourth president of the United States might not have been the forty-fourth president of the United States (for example, if Obama had lost to Clinton), but it is not true that Barack Obama might not have been Barack Obama. (Obama might not have been called "Barack Obama," but that is another matter.) This shows that a contingent identity, "X = Y," is possible only if either of the two expressions, "X" or "Y," is a nonrigid designator, an expression that can refer to different things in different worlds.

Consider the term "C-fiber excitation": Could this designator be nonrigid? It would seem not: How could an event that in fact is the excitation of C-fibers not have been one? How could an event that is an instance of C-fiber excitation be, say, a volcano eruption or a collision of two stars in another possible world? A world in which no C-fiber excitation ever occurs is a world in which this event, which is a C-fiber excitation, does not occur. The term "pain" also

seems rigid. If you are inclined to take the painfulness of pain as its essential defining property, you will say that "pain" rigidly designates an event or state with this quality of painfulness and that the expression designates an event of that sort across all possible worlds. A world in which nothing ever hurts is a world without pain.

It follows that if pain = Cfs, then this must be a necessary truth—that is, it must hold in every possible world. Descartes famously claimed that it is possible for him to exist as a thinking and conscious thing even without a body. If that is possible, then pain could exist even if Cfs did not. Some philosophers have argued that "zombies"—creatures that are physically just like us but have no consciousness—are possible; that is, there are possible worlds inhabited by zombies. If so, Cfs could exist without being accompanied by pain. If these are real possibilities, then "pain = Cfs" cannot be a necessary truth. Then, by the principle that if X and Y are rigid designators, the identity "X = Y" is necessarily true, if true, it follows that "pain = Cfs" is false—false in this world. More generally, psychoneural identities are all false.

Many mind-brain identity theorists would be likely to dispute the claim that it is possible that pain can exist even if Cfs does not, and they would question the claim that zombies are a real possibility. We can grant, they will argue, that in some sense these situations are "conceivable," that we can "imagine" such possibilities. But the fact that a situation is conceivable or imaginable does not entail that it is genuinely possible. For example, it is conceivable, they will say, that water is not H_2O and that heat is not molecular kinetic energy; the concept of water and the concept of H_2O are logically unrelated to each other, and there is no conceptual incoherence or contradiction in the thought that water ≠ H_2O. And we might even say that "water ≠ H_2O" is *epistemically possible*: For all that people knew about water and other things not so long ago, it was possible that water could have turned out to be something other than H_2O. That is, for all we knew a couple hundred years ago, we might be living on a planet with XYZ, rather than H_2O, coming out of the tap, filling our lakes and rivers, and so on, where XYZ is observationally indistinguishable from H_2O, although wholly different in molecular structure. Nonetheless, water = H_2O, and necessarily so. The gist of the reply by the identity theorists then is that conceivability does not entail real metaphysical possibility and that this is shown by a posteriori necessary identities like "water = H_2O" and "heat = molecular kinetic energy." For them, psychoneural identities, "pain = Cfs" and the like, are necessary a posteriori truths just like these scientific identities. Issues about conceivability and possibility are highly complex and contentious, and they are being actively debated, without a consensus resolution in sight.[21]

The Multiple Realization Argument

The psychoneural identity theory says that pain is C-fiber excitation. But that implies that unless an organism has C-fibers, it cannot have pain. But aren't there pain-capable organisms, like reptiles and mollusks, with nervous systems very different from the human nervous system? Perhaps in these species the cells that work as nociceptive neurons—pain-receptor neurons—are not like human C-fibers at all; how can we be sure that all pain-capable animals have C-fibers? Can the identity physicalist reply that it should be possible to come up with a more abstract and general physiological description of a brain state common to all organisms, across all species, that are in pain? This seems highly unlikely, and in any case, how about inorganic systems? Could there not be intelligent extraterrestrial creatures with a complex and rich mental life but whose biology is not carbon-based? And is it not conceivable—in fact, nomologically possible if not practically feasible—to build intelligent electromechanical robots to which we would be willing to attribute various mental states (perceptual and cognitive states, if not sensations and emotions)? Moreover, the neural substrates of highly specific mental states (e.g., having the belief that winters are colder in New Hampshire than in Rhode Island) can differ from person to person and may change over time even in a single person through maturation, learning, and brain injuries. Does it make sense to think that some single neural state is shared by all persons who believe that cats are smarter than dogs, or that 7 + 5 = 12? Moreover, we should keep in mind that if pain is identical with some physical state, this must hold not only in actual organisms and systems but in all possible organisms and systems. This is so because, as we saw earlier in our discussion of the modal argument, such identities, if true, must be necessarily true.

These considerations are widely thought to show that any mental state is "multiply realizable"[22] in a large variety of physical-biological systems, with the consequence that it is not possible to identify mental states with physical states. If pain is identical with a physical state, it must be identical with some *particular* physical state, but there is no single neural correlate or substrate of pain. On the contrary, there must be indefinitely many physical states that can "realize" (or "instantiate," or "implement") pain in all sorts of pain-capable organisms and systems. So pain, as a type of mental state, cannot be identified with a neural state type or with any other physical state type.

This is the influential and widely known "multiple realization argument" that Hilary Putnam and others advanced in the late 1960s and early 1970s. It has had a critical impact on the way philosophy of mind has developed since

then. It was this argument, rather than any of the other difficulties, that brought about an unexpectedly early decline of psychoneural identity theory. What made the multiple realization argument distinctive, and different from other sundry objections, was that it brought with it a fresh and original conception of the mental, which offered an attractive alternative approach to the nature of mind. This is functionalism, still the reigning orthodoxy on the nature of mentality and the status of psychology. We turn to this influential view in the next two chapters.

Reductive and Nonreductive Physicalism

The psychoneural identity theory, or identity physicalism, is a form of reductive physicalism. It reductively identifies mental states with neural states of the brain. It is also called type physicalism, since it identifies types, or kinds, of mental states, like pain, thirst, anger, and so on, with types and kinds of neural-physical states. That is, psychological types, or properties, are claimed to be identical with neural-physical types and properties. Thus, type physicalism contrasts with the so-called token physicalism (see chapter 1), according to which, though psychological types and properties are not neural-physical types, each individual, "token" psychological event, like this particular pain I am experiencing now, is in fact a neural event. This means that different tokens, or instances, of a single mental kind may, and usually will, fall under distinct neural kinds. Both you and an octopus experience a pain, but your pain is an instance of C-fiber stimulation and the octopus pain is an instance of (let's say) O-fiber stimulation. As you can tell, token physicalism is inspired by considerations of multiple realization of psychological states.

Since the 1970s, chiefly on account of the influence of the multiple realization argument, reductive physicalism has had a rough time of it, although of late it has shown renewed strength and signs of a revival. As reductionism's fortunes declined, nonreductive physicalism (see chapter 1) rapidly gained strength and influence, and it has reigned as the dominant and virtually unchallenged position on the mind-body problem for the past several decades. This is the view that mental properties, along with other "higher-level" properties of the special sciences, like biology, geology, and the social sciences, resist reduction to the basic physical domain. An antireductionist view of this kind has also served as an influential philosophical foundation of psychology and cognitive science, providing support for the claim that these sciences are autonomous, each with its own distinctive methodology and system of concepts and not answerable to the

methodological or explanatory constraints of more fundamental sciences. Thus, the most widely accepted form of physicalism today combines substance physicalism with property dualism: All concrete individual things in this world are physical, but complex physical systems can, and sometimes do, exhibit properties that are not reducible to "lower-level" physical properties. Among these irreducible properties are, most notably, psychological properties, including those investigated in the psychological and cognitive sciences.

But nonreductive physicalism, above all, is a form of physicalism. What makes it physicalistic? In what do its credentials as physicalism consist? Part of the answer is that it accepts substance physicalism. It rejects Cartesian mental substances and other supposed nonphysical things in space-time, and of course there is nothing outside space-time. Although the nonreductive physicalist denies the physical reducibility of the mental, she nonetheless accepts a close and intimate relationship between mental properties and physical properties, and this is mind-body supervenience (see chapter 1). We may call this *supervenience physicalism*. Some nonreductive physicalists will go a step further and maintain that their irreducible mental properties are "physically realized" or "physically implemented." This is the so-called *realization physicalism*.[23] There will be more in the next chapter on the idea of physical realization; here, the point to note is that the realization relation is stronger than supervenience, and hence that realization physicalism is a stronger thesis than supervenience physicalism. If mind-body realization holds, then mind-body supervenience holds, but not the other way around.

In any case, in committing herself to the supervenience, or realization, relation between mental and physical properties, the nonreductive physicalist goes beyond mere property dualism. It should be clear that property dualism as such does not require the thesis that the mental character of a being is dependent on, or determined by, its physical nature, as mind-body supervenience requires, or that mental properties are physically realized (if they are realized at all). Mental properties, though instantiated in physical systems, might yet be independent of their physical properties.

Moreover, nonreductive physicalists are mental realists who believe in the reality of mental properties; they regard mental properties as genuine properties the possession of which makes a difference—a causal difference. Part of the belief in the reality of mental properties is to believe in their causal efficacy. An organism, in virtue of having a mental property (say, wanting a drink of water or being in pain), acquires powers and propensities to act or be acted upon in certain ways. Summarizing all this, nonreductive physicalism, as standardly understood, comprises the following four claims:

Substance Physicalism. The space-time world consists exclusively of bits of matter and their aggregates.

Irreducibility of the Mental. Mental properties are not reducible to physical properties.

Mind-Body Supervenience or *Realization*. Either (a) mental properties supervene on physical properties, or (b) mental properties, when they are realized, are realized by physical properties.

Mental Causal Efficacy. Mental properties are causally efficacious; mental events are sometimes causes of other events, both physical and mental.

Nonreductive physicalism, understood as the conjunction of these four theses, has been the most influential position on the mental-physical relation. We can think of property dualism as the conjunction of the first, second, and fourth doctrines—that is, all but mind-body supervenience/realization. Besides its acceptance of substance physicalism, what makes nonreductive physicalism a serious physicalism is its commitment to mind-body supervenience/realization. Property dualism that rejects mind-body supervenience/realization seems, prima facie, to be a possible position; however, this form of property dualism has not found strong advocates and remains largely undeveloped. And there may be a good reason for this: In rejecting supervenience/realization, you take the mental as constituting its own realm separate from the physical and it is difficult to see how you would be able to explain the causal efficacy of the mental in the physical world. You might very well run into troubles of the kind Descartes had in explaining how immaterial minds could causally interact with material things (see chapter 2). The rejection of mind-body supervenience, therefore, may force you to give up mental causal efficacy, and this is not an option for most (see chapter 7).

In accepting the irreducibility thesis, nonreductive physicalism attempts to honor the special position that thought and consciousness enjoy in our conception of ourselves among the things of this world. As was noted above, the irreducibility thesis is also an affirmation of the autonomy of psychology and cognitive science as sciences in their own right, not constrained by more basic sciences. In accepting the causal efficacy of the mental, the nonreductive physicalist not only acknowledges what seems so familiar and obvious to common sense, but at the same time, it declares psychology and cognitive science to be

genuine sciences capable of generating law-based causal explanations and predictions. All in all, it is an attractive package, and it is not difficult to understand its appeal and staying power.

However, all that may only be wishful thinking. The story may be too good to be true. There have recently been significant objections and criticisms of the nonreductive aspect of nonreductive physicalism, and these collectively have generated enough pressure for many philosophers to reconsider its viability. We will see some of the difficulties nonreductive physicalism faces in regard to mental causation later (see chapter 7).

For Further Reading

The classic sources of the mind-brain identity theory are Herbert Feigl, "The 'Mental' and the 'Physical,'" and J. J. C. Smart, "Sensations and Brain Processes," both of which are available in *Philosophy of Mind: Classical and Contemporary Readings*, edited by David J. Chalmers. The Smart article is widely reprinted in anthologies on philosophy of mind. For more recent book-length treatments of physicalism and related issues, see Christopher S. Hill, *Sensations: A Defense of Type Physicalism*; Jeffrey Poland, *Physicalism: The Philosophical Foundation*; Andrew Melnyk, *A Physicalist Manifesto*; Thomas W. Polger, *Natural Minds*; Jaegwon Kim, *Physicalism, or Something Near Enough*; Daniel Stoljar, *Physicalism*.

For criticisms, see Saul Kripke, *Naming and Necessity*, lecture 3. John Heil's anthology, *Philosophy of Mind: A Guide and Anthology*, includes three essays (by John Foster, Peter Forrest, and E. J. Lowe) that are worth examining in a section with the title "Challenges to Contemporary Materialism." A very recent collection of critical essays on physicalism is *The Waning of Materialism*, edited by Robert C. Koons and George Bealer.

For the multiple realization argument against the psychoneural identity theory, the original sources are Hilary Putnam, "Psychological Predicates," later retitled "The Nature of Mental States," and Jerry Fodor's "Special Sciences, or the Disunity of Science as a Working Hypothesis." For recent reevaluations of the argument, see Jaegwon Kim, "Multiple Realization and the Metaphysics of Reduction"; William Bechtel and Jennifer Mundale, "Multiple Realizability Revisited: Linking Cognitive and Neural States." There is an extensive discussion of realization and multiple realizability in Lawrence Shapiro, *The Mind Incarnate*.

On the status of nonreductive physicalism, see Kim, "The Myth of Nonreductive Physicalism" and "Multiple Realization and the Metaphysics of Reduction"; Andrew Melnyk, "Can Physicalism Be Non-Reductive?" For responses:

Ned Block, "Anti-Reductionism Slaps Back"; Jerry Fodor, "Special Sciences: Still Autonomous After All These Years"; Louise Antony, "Everybody Has Got It: A Defense of Non-Reductive Materialism."

Notes

1. See René Descartes, *The Passions of the Soul*.

2. See Thomas H. Huxley, "On the Hypothesis That Animals Are Automata, and Its History."

3. Samuel Alexander, *Space, Time, and Deity*. Vol. 2, p. 47. "Natural piety" is an expression made famous by the poet William Wordsworth.

4. J. J. C. Smart, "Sensations and Brain Processes." U. T. Place's "Is Consciousness a Brain Process?" published in 1956, predates Smart's article as perhaps the first modern statement of the identity theory.

5. J. J. C. Smart, "Sensations and Brain Processes," p. 117 (in the reprint version in *Philosophy of Mind: A Guide and Anthology*, ed. John Heil. Emphasis in the original).

6. See the entry "William of Ockham" in the *Macmillan Encyclopedia of Philosophy*, 2nd ed.

7. Herbert Feigl, "The 'Mental' and the 'Physical,'" p. 428.

8. See Gilbert Harman, "The Inference to the Best Explanation." For a critique of the principle, see Bas Van Fraassen, *Laws and Symmetry*.

9. This example comes from Christopher S. Hill, *Sensations: A Defense of Type Materialism*, p. 24. Hill's book includes an extremely clear and forceful presentation of explanatory argument I.

10. This is substantially the form in which Brian McLaughlin formulates his explanatory argument. See his "In Defense of New Wave Materialism: A Response to Horgan and Tienson." Hill (see note 10) and McLaughlin are two leading proponents of this form of the explanatory argument. However, McLaughlin does not explicitly invoke the rule of inference to the best explanation. See also Andrew Melnyk, *A Physicalist Manifesto*.

11. J. J. C. Smart, "Sensations and Brain Processes," p. 126.

12. Ned Block and Robert Stalnaker, "Conceptual Analysis, Dualism, and the Explanatory Gap," p. 24.

13. It is debatable whether it really is improper, or wrongheaded (as Block and Stalnaker put it), to ask for explanations of psychoneural correlations. One might argue that such explanatory demands are perfectly in order, and that to the extent that physicalism is unable to meet them, it is a limited and flawed doctrine.

14. For further discussion, see David Papineau, "The Rise of Physicalism" and *Thinking About Consciousness,* chapter 1.

15. You might recall the pairing problem discussed in chapter 2, in connection with Descartes's interactionist dualism.

16. There is another issue with the argument as presented, which is discussed in chapter 7 on mental causation; see the section on the "exclusion argument."

17. This objection is worked out in detail in Jerome Shaffer, "Mental Events and the Brain." The original form of this argument is credited to Max Black by J. J. C. Smart in his "Sensations and Brain Processes."

18. See J. J. C. Smart, "Sensations and Brain Processes."

19. This strategy originated in Brian Loar, "Phenomenal States." For more recent discussions, see *Phenomenal Concepts and Phenomenal Knowledge*, ed. Torin Alter and Sven Walter. Also helpful are: Katalin Balog, "Phenomenal Concepts," in *Oxford Handbook of Philosophy of Mind*, ed. Brian McLaughlin et al.; Peter Carruthers and Benedicte Veillet, "The Phenomenal Concept Strategy."

20. This neo-Cartesian modal argument is due to Saul Kripke. See his *Naming and Necessity*, especially lecture 3, in which the argument is presented in detail.

21. For further discussion of these issues, see the essays in *Conceivability and Possibility*, edited by Tamar Szabo Gendler and John Hawthorne.

22. The terms "variably realizable" and "variable realization" are commonly used by British writers.

23. I believe Andrew Melnyk first used this term in his *A Physicalist Manifesto*. Jaegwon Kim used the term "physical realizationism" earlier in *Mind in a Physical World*, but "realization physicalism" is better.