Chapter 5

Reductionism and the Irreducibility of Consciousness

The view of the relation between mind and body that I have been putting forward is sometimes called "reductionist," sometimes "antireductionist." It is often called "emergentism," and is generally regarded as a form of "supervenience." I am not sure that any one of these attributions is at all clear, but a number of issues surround these mysterious terms, and in this chapter I will explore some of them.

I. Emergent Properties

Suppose we have a system, *S*, made up of elements *a*, *b*, *c*. . . For example, *S* might be a stone and the elements might be molecules. In general, there will be features of *S* that are not, or not necessarily, features of *a*, *b*, *c*. . . For example, *S* might weigh ten pounds, but the molecules individually do not weigh ten pounds. Let us call such features "system features." The shape and the weight of the stone are system features. Some system features can be deduced or figured out or calculated from the features of *a*, *b*, *c*. . . just from the way these are composed and arranged (and sometimes from their relations to the rest of the environment). Examples of these would be shape, weight, and velocity. But some other system features cannot be figured out just from the composition of the elements and environmental relations; they have to be explained in terms of the causal interactions among the elements. Let's call these "causally emergent system features." Solidity, liquidity, and transparency are examples of causally emergent system features.

On these definitions, consciousness is a causally emergent property of systems. It is an emergent feature of certain systems of neurons in the same way that solidity and liquidity are emergent features of systems of molecules. The existence of consciousness can be explained by the causal interactions between elements of the brain at the micro level, but consciousness cannot itself be deduced or calculated from the sheer physical structure of the neurons without some additional account of the causal relations between them.

This conception of causal emergence, call it "emergent1," has to be distinguished from a much more adventurous conception, call it "emergent2." A feature *F* is emergent2 iff *F* is emergent1 and *F* has causal powers that cannot be explained by the causal interactions of *a*, *b*, *c*. . . If consciousness were emergent2, then consciousness could cause things that could not be explained by the causal behavior of the neurons. The naive idea here is that consciousness gets squirted out by the behavior of the neurons in the brain, but once it has been squirted out, it then has a life of its own.

It should be obvious from the previous chapter that on my view consciousness is emergent1, but not emergent2. In fact, I cannot think of anything that is emergent2, and it seems unlikely that we will be able to find any features that are emergent2, because the existence of any such features would seem to violate even the weakest principle of the transitivity of causation.

II. Reductionism

Most discussions of reductionism are extremely confusing. Reductionism as an ideal seems to have been a feature of positivist philosophy of science, a philosophy now in many respects discredited. However, discussions of reductionism still survive, and the basic intuition that underlies the concept of reductionism seems to be the idea that certain things might be shown to be *nothing but* certain other sorts of things. Reductionism, then, leads to a peculiar form of the identity relation

that we might as well call the "nothing-but" relation: in general, *A*'s can be reduced to *B*'s, iff *A*'s are nothing but *B*'s.

However, even within the nothing-but relation, people mean so many different things by the notion of "reduction" that we need to begin by making several distinctions. At the very outset it is important to be clear about what the relata of the relation are. What is its domain supposed to be: objects, properties, theories, or what? I find at least five different senses of "reduction"—or perhaps I should say five different kinds of reduction—in the theoretical literature, and I want to mention each of them so that we can see which are relevant to our discussion of the mind-body problem.

1. Ontological Reduction

The most important form of reduction is ontological reduction. It is the form in which objects of certain types can be shown to consist in nothing but objects of other types. For example, chairs are shown to be nothing but collections of molecules. This form is clearly important in the history of science. For example, material objects in general can be shown to be nothing but collections of molecules, genes can be shown to consist in nothing but DNA molecules. It seems to me this form of reduction is what the other forms are aiming at.

2. Property Ontological Reduction

This is a form of ontological reduction, but it concerns properties. For example, heat (of a gas) is nothing but the mean kinetic energy of molecule movements. Property reductions for properties corresponding to theoretical terms, such as "heat," "light," etc., are often a result of theoretical reductions.

3. Theoretical Reduction

Theoretical reductions are the favorite of theorists in the literature, but they seem to me rather rare in the actual practice of science, and it is perhaps not surprising that the same half dozen examples are given over and over in the standard textbooks. From the point of view of scientific explanation,

theoretical reductions are mostly interesting if they enable us to carry out ontological reductions. In any case, theoretical reduction is primarily a relation between theories, where the laws of the reduced theory can (more or less) be deduced from the laws of the reducing theory. This demonstrates that the reduced theory is nothing but a special case of the reducing theory. The classical example that is usually given in textbooks is the reduction of the gas laws to the laws of statistical thermodynamics.

4. Logical or Definitional Reduction

This form of reduction used to be a great favorite among philosophers, but in recent decades it has fallen out of fashion. It is a relation between words and sentences, where words and sentences referring to one type of entity can be translated without any residue into those referring to another type of entity. For example, sentences about the average plumber in Berkeley are reducible to sentences about specific individual plumbers in Berkeley; sentences about numbers, according to one theory, can be translated into, and hence are reducible to, sentences about sets. Since the words and sentences are *logically* or *definitionally* reducible, the corresponding entities referred to by the words and sentences are *ontologically* reducible. For example, numbers are nothing but sets of sets.

5. Causal Reduction

This is a relation between any two types of things that can have causal powers, where the existence and a fortiori the causal powers of the reduced entity are shown to be entirely explainable in terms of the causal powers of the reducing phenomena. Thus, for example, some objects are solid and this has causal consequences: solid objects are impenetrable by other objects, they are resistant to pressure, etc. But these causal powers can be causally explained by the causal powers of vibratory movements of molecules in lattice structures.

Now when the views I have urged are accused of being reductionist—or sometimes insufficiently reductionist—which

of these various senses do the accusers have in mind? I think that theoretical reduction and logical reduction are not intended. Apparently the question is whether the causal reductionism of my view leads—or fails to lead—to ontological reduction. I hold a view of mind/brain relations that is a form of causal reduction, as I have defined the notion: Mental features are caused by neurobiological processes. Does this imply ontological reduction?

In general in the history of science, successful causal reductions tend to lead to ontological reductions. Because where we have a successful causal reduction, we simply redefine the expression that denotes the reduced phenomena in such a way that the phenomena in question can now be identified with their causes. Thus, for example, color terms were once (tacitly) defined in terms of the subjective experience of color perceivers; for example, "red" was defined ostensively by pointing to examples, and then real red was defined as whatever seemed red to "normal" observers under "normal" conditions. But once we have a causal reduction of color phenomena to light reflectances, then, according to many thinkers, it becomes possible to redefine color expressions in terms of light reflectances. We thus carve off and eliminate the subjective experience of color from the "real" color. Real color has undergone a property ontological reduction to light reflectances. Similar remarks could be made about the reduction of heat to molecular motion, the reduction of solidity to molecular movements in lattice structures, and the reduction of sound to air waves. In each case, the causal reduction leads naturally to an ontological reduction by way of a redefinition of the expression that names the reduced phenomenon. Thus, to continue with the example of "red," once we know that the color experiences are caused by a certain sort of photon emission, we then redefine the word in terms of the specific features of the photon emission. "Red," according to some theorists, now refers to photon emissions of 600 nanometers. It thus follows trivially that the color red is nothing but photon emissions of 600 nanometers.

The general principle in such cases appears to be this: Once a property is seen to be *emergent1*, we automatically get a causal reduction, and that leads to an ontological reduction, by redefinition if necessary. The general trend in ontological reductions that have a scientific basis is toward greater generality, objectivity, and redefinition in terms of underlying causation.

So far so good. But now we come to an apparently shocking asymmetry. When we come to consciousness, we cannot perform the ontological reduction. Consciousness is a causally emergent property of the behavior of neurons, and so consciousness is causally reducible to the brain processes. But—and this is what seems so shocking—a perfect science of the brain would still not lead to an ontological reduction of consciousness in the way that our present science can reduce heat, solidity, color, or sound. It seems to many people whose opinions I respect that the irreducibility of consciousness is a primary reason why the mind-body problem continues to seem so intractable. Dualists treat the irreducibility of consciousness as incontrovertible proof of the truth of dualism. Materialists insist that consciousness must be reducible to material reality, and that the price of denying the reducibility of consciousness would be the abandonment of our overall scientific world view.

I will briefly discuss two questions: First, I want to show why consciousness is irreducible, and second, I want to show why it does not make any difference at all to our scientific world view that it should be irreducible. It does not force us to property dualism or anything of the sort. It is a trivial consequence of certain more general phenomena.

III. Why Consciousness Is an Irreducible Feature of Physical Reality

There is a standard argument to show that consciousness is not reducible in the way that heat, etc., are. In different ways the argument occurs in the work of Thomas Nagel (1974), Saul

Kripke (1971), and Frank Jackson (1982). I think the argument is decisive, though it is frequently misunderstood in ways that treat it as merely epistemic and not ontological. It is sometimes treated as an epistemic argument to the effect that, for example, the sort of third-person, objective knowledge we might possibly have of a bat's neurophysiology would still not include the first-person, subjective experience of what it feels like to be a bat. But for our present purposes, the point of the argument is ontological and not epistemic. It is a point about what real features exist in the world and not, except derivatively, about how we know about those features.

Here is how it goes: Consider what facts in the world make it the case that you are now in a certain conscious state such as pain. What fact in the world corresponds to your true statement, "I am now in pain"? Naively, there seem to be at least two sorts of facts. First and most important, there is the fact that you are now having certain unpleasant conscious sensations, and you are experiencing these sensations from your subjective, first-person point of view. It is these sensations that are constitutive of your present pain. But the pain is also caused by certain underlying neurophysiological processes consisting in large part of patterns of neuron firing in your thalamus and other regions of your brain. Now suppose we tried to reduce the subjective, conscious, first-person sensation of pain to the objective, third-person patterns of neuron firings. Suppose we tried to say the pain is really "nothing but" the patterns of neuron firings. Well, if we tried such an ontological reduction, the essential features of the pain would be left out. No description of the third-person, objective, physiological facts would convey the subjective, first-person character of the pain, simply because the first-person features are different from the third-person features. Nagel states this point by contrasting the objectivity of the third-person features with the what-it-is-like features of the subjective states of consciousness. Jackson states the same point by calling attention to the fact that someone who had a complete knowledge of the neurophysiology of a mental phenomenon such as pain would still

not know what a pain was if he or she did not know what it felt like. Kripke makes the same point when he says that pains could not be identical with neurophysiological states such as neuron firings in the thalamus and elsewhere, because any such identity would have to be necessary, because both sides of the identity statement are rigid designators, and yet we know that the identity could not be necessary.[1] This fact has obvious epistemic consequences: my knowledge that I am in pain has a different sort of basis than my knowledge that you are in pain. But the antireductionist point of the argument is ontological and not epistemic.

So much for the antireductionist argument. It is ludicrously simple and quite decisive. An enormous amount of ink has been shed trying to answer it, but the answers are all so much wasted ink. But to many people it seems that such an argument paints us into a corner. To them it seems that if we accept that argument, we have abandoned our scientific world view and adopted property dualism. Indeed, they would ask, what is property dualism but the view that there are irreducible mental properties? In fact, doesn't Nagel accept property dualism and Jackson reject physicalism precisely because of this argument? And what is the point of scientific reductionism if it stops at the very door of the mind? So I now turn to the main point of this discussion.

IV. Why the Irreducibility of Consciousness Has No Deep Consequences

To understand fully why consciousness is irreducible, we have to consider in a little more detail the pattern of reduction that we found for perceivable properties such as heat, sound, color, solidity, liquidity, etc., and we have to show how the attempt to reduce consciousness differs from the other cases. In every case the ontological reduction was based on a prior causal reduction. We discovered that a surface feature of a phenomenon was caused by the behavior of the elements of an underlying microstructure. This is true both in the cases in

which the reduced phenomenon was a matter of subjective appearances, such as the "secondary qualities" of heat or color; and in the cases of the "primary qualities" such as solidity, in which there was both an element of subjective appearance (solid things feel solid), and also many features independent of subjective appearances (solid things, e.g., are resistant to pressure and impenetrable by other solid objects). But in each case, for both the primary and secondary qualities, the point of the reduction was to carve off the surface features and redefine the original notion in terms of the causes that produce those surface features.

Thus, where the surface feature is a subjective appearance, we redefine the original notion in such a way as to exclude the appearance from its definition. For example, pretheoretically our notion of heat has something to do with perceived temperatures: Other things being equal, hot is what feels hot to us, cold is what feels cold. Similarly with colors: Red is what looks red to normal observers under normal conditions. But when we have a theory of what causes these and other phenomena, we discover that it is molecular movements causing sensations of heat and cold (as well as other phenomena such as increases in pressure), and light reflectances causing visual experiences of certain sorts (as well as other phenomena such as movements of light meters). We then *redefine* heat and color in terms of the underlying causes of both the subjective experiences and the other surface phenomena. And in the redefinition we eliminate any reference to the subjective appearances and other surface effects of the underlying causes. "Real" heat is now defined in terms of the kinetic energy of the molecular movements, and the subjective feel of heat that we get when we touch a hot object is now treated as just a subjective appearance caused by heat, as an effect of heat. It is no longer part of real heat. A similar distinction is made between real color and the subjective experience of color. The same pattern works for the primary qualities: Solidity is defined in terms of the vibratory movements of molecules in lattice structures, and objective, observer-independent features, such as

impenetrability by other objects, are now seen as surface effects of the underlying reality. Such redefinitions are achieved by way of carving off all of the surface features of the phenomenon, whether subjective or objective, and treating them as effects of the real thing.

But now notice: The actual pattern of the facts in the world that correspond to statements about particular forms of heat such as specific temperatures are quite similar to the pattern of facts in the world that correspond to statements about particular forms of consciousness, such as pain. If I now say, "It's hot in this room," what are the facts? Well, first there is a set of "physical" facts involving the movement of molecules, and second there is a set of "mental" facts involving my subjective experience of heat, as caused by the impact of the moving air molecules on my nervous system. But similarly with pain. If I now say , "I am in pain," what are the facts? Well, first there is a set of "physical" facts involving my thalamus and other regions of the brain, and second there is a set of "mental" facts involving my subjective experience of pain. So why do we regard heat as reducible and pain as irreducible? The answer is that what interests us about heat is not the subjective appearance but the underlying physical causes. Once we get a causal reduction, we simply redefine the notion to enable us to get an ontological reduction. Once you know all the facts about heat—facts about molecule movements, impact on sensory nerve endings, subjective feelings, etc.—the reduction of heat to molecule movements involves no new *fact* whatever. It is simply a trivial consequence of the redefinition. We don't first discover all the facts and then discover a new fact, the fact that heat is reducible; rather, we simply redefine heat so that the reduction follows from the definition. But this redefinition does not eliminate, and was not intended to eliminate, the subjective experiences of heat (or color, etc.) from the world. They exist the same as ever.

We might not have made the redefinition. Bishop Berkeley, for example, refused to accept such redefinitions. But it is easy to see why it is rational to make such redefinitions and accept

their consequences: To get a greater understanding and control of reality, we want to know how it works causally, and we want our concepts to fit nature at its causal joints. We simply redefine phenomena with surface features in terms of the underlying causes. It then looks like a new discovery that heat is *nothing but* mean kinetic energy of molecule movement, and that if all subjective experiences disappeared from the world, real heat would still remain. But this is not a new discovery, it is a trivial consequence of a new definition. Such reductions do not show that heat, solidity, etc., do not really exist in the way that, for example, new knowledge showed that mermaids and unicorns do not exist.

Couldn't we say the same thing about consciousness? In the case of consciousness, we do have the distinction between the "physical" processes and the subjective "mental" experiences, so why can't consciousness be redefined in terms of the neurophysiological processes in the way that we redefined heat in terms of underlying physical processes? Well, of course, if we insisted on making the redefinition, we could. We could simply define, for example, "pain" as patterns of neuronal activity that cause subjective sensations of pain. And if such a redefinition took place, we would have achieved the same sort of reduction for pain that we have for heat. But of course, the reduction of pain to its physical reality still leaves the subjective experience of pain unreduced, just as the reduction of heat left the subjective experience of heat unreduced. Part of the point of the reductions was to carve off the subjective experiences and exclude them from the definition of the real phenomena, which are now defined in terms of those features that interest us most. But where the phenomena that interest us most are the subjective experiences themselves, there is no way to carve anything off. Part of the point of the reduction in the case of heat was to distinguish between the subjective appearance on the one hand and the underlying physical reality on the other. Indeed, it is a general feature of such reductions that the phenomenon is defined in terms of the "reality" and not in terms of the "appearance." But we can't make that

sort of appearance-reality distinction for consciousness because consciousness consists in the appearances themselves. *Where appearance is concerned we cannot make the appearance-reality distinction because the appearance is the reality.*

For our present purposes, we can summarize this point by saying that consciousness is not reducible in the way that other phenomena are reducible, not because the pattern of facts in the real world involves anything special, but because the reduction of other phenomena depended in part on distinguishing between "objective physical reality," on the one hand, and mere "subjective appearance," on the other; and eliminating the appearance from the phenomena that have been reduced. But in the case of consciousness, its reality is the appearance; hence, the point of the reduction would be lost if we tried to carve off the appearance and simply defined consciousness in terms of the underlying physical reality. In general, the pattern of our reductions rests on rejecting the subjective epistemic basis for the presence of a property as part of the ultimate constituent of that property. We find out about heat or light by feeling and seeing, but we then define the phenomenon in a way that is independent of the epistemology. Consciousness is an exception to this pattern for a trivial reason. The reason, to repeat, is that the reductions that leave out the epistemic bases, the appearances, cannot work for the epistemic bases themselves. In such cases, the appearance is the reality.

But this shows that the irreducibility of consciousness is a trivial consequence of the pragmatics of our definitional practices. A trivial result such as this has only trivial consequences. It has no deep metaphysical consequences for the unity of our overall scientific world view. It does not show that consciousness is not part of the ultimate furniture of reality or cannot be a subject of scientific investigation or cannot be brought into our overall physical conception of the universe; it merely shows that in the way that we have decided to carry out reductions, consciousness, by definition, is excluded from a certain pattern of reduction. Consciousness fails to be reducible, not

because of some mysterious feature, but simply because by definition it falls outside the pattern of reduction that we have chosen to use for pragmatic reasons. Pretheoretically, consciousness, like solidity, is a surface feature of certain physical systems. But unlike solidity, consciousness cannot be redefined in terms of an underlying microstructure, and the surface features then treated as mere effects of real consciousness, without losing the point of having the concept of consciousness in the first place.

So far, the argument of this chapter has been conducted, so to speak, from the point of view of the materialist. We can summarize the point I have been making as follows: The contrast between the reducibility of heat, color, solidity, etc., on the one hand, and the irreducibility of conscious states, on the other hand, does not reflect any distinction in the structure of reality, but a distinction in our definitional practices. We could put the same point from the point of view of the property dualist as follows: The apparent contrast between the irreducibility of consciousness and the reducibility of color, heat, solidity, etc., really was *only* apparent. We did not really eliminate the subjectivity of red, for example, when we reduced red to light reflectances; we simply stopped calling the subjective part "red." We did not eliminate any subjective phenomena whatever with these "reductions"; we simply stopped calling them by their old names. Whether we treat the irreducibility from the materialist or from the dualist point of view, we are still left with a universe that contains an irreducibly subjective physical component as a component of physical reality.

To conclude this part of the discussion, I want to make clear what I am saying and what I am not saying. I am not saying that consciousness is not a strange and wonderful phenomenon. I think, on the contrary, that we ought to be amazed by the fact that evolutionary processes produced nervous systems capable of causing and sustaining subjective conscious states. As I remarked in chapter 4, consciousness is as empirically mysterious to us now as electromagnetism was previously, when people thought the universe must operate

entirely on Newtonian principles. But I am saying that once the existence of (subjective, qualitative) consciousness is granted (and no sane person can deny its existence, though many pretend to do so), then there is nothing strange, wonderful, or mysterious about its *irreducibility*. Given its existence, its irreducibility is a trivial consequence of our definitional practices. Its irreducibility has no untoward scientific consequences whatever. Furthermore, when I speak of the irreducibility of consciousness, I am speaking of its *irreducibility according to standard patterns of reduction*. No one can rule out a priori the possibility of a major intellectual revolution that would give us a new—and at present unimaginable—conception of reduction, according to which consciousness would be reducible.

V. Supervenience

In recent years there has been a lot of heavy going about a relationship between properties called "supervenience" (e.g., Kim 1979, 1982; Haugeland 1982). It is frequently said in discussions in the philosophy of mind that the mental is supervenient on the physical. Intuitively, what is meant by this claim is that mental states are totally dependent on corresponding neurophysiological states in the sense that a difference in mental states would necessarily involve a corresponding difference in neurophysiological states. If, for example, I go from a state of being thirsty to a state of no longer being thirsty, then there must have been some change in my brain states corresponding to the change in my mental states.

On the account that I have been proposing, mental states are supervenient on neurophysiological states in the following respect: Type-identical neurophysiological causes would have type-identical mentalistic effects. Thus, to take the famous brain-in-the-vat example, if you had two brains that were type-identical down to the last molecule, then the causal basis of the mental would guarantee that they would have the same mental phenomena. On this characterization of the supervenience relation, the supervenience of the mental on the physical is marked by the fact that physical states are causally sufficient, though not necessarily causally necessary, for the corresponding mental states. That is just another way of saying that as far as this definition of supervenience is concerned, sameness of neurophysiology guarantees sameness of mentality; but sameness of mentality does not guarantee sameness of neurophysiology.

It is worth emphasizing that this sort of supervenience is *causal* supervenience. Discussions of supervenience were originally introduced in connection with ethics, and the notion in question was not a causal notion. In the early writings of Moore (1922) and Hare (1952), the idea was that moral properties are supervenient on natural properties, that two objects cannot differ solely with respect to, for example, their goodness. If one object is better than another, there must be some other feature in virtue of which the former is better than the latter. But this notion of moral supervenience is not a causal notion. That is, the features of an object that make it good do not *cause* it to be good, they rather *constitute* its goodness. But in the case of mind/brain supervenience, the neural phenomena cause the mental phenomena.

So there are at least two notions of supervenience: a constitutive notion and a causal notion. I believe that only the causal notion is important for discussions of the mind-body problem. In this respect my account differs from the usual accounts of the supervenience of the mental on the physical. Thus Kim (1979, especially p. 45ff.) claims that we should not think of the relation of neural events to their supervening mental events as causal, and indeed he claims that supervening mental events have no causal status apart from their supervenience on neurophysiological events that have "a more direct causal role." "If this be epiphenomenalism, let us make the most of it," he says cheerfully (p. 47).

I disagree with both of these claims. It seems to me obvious from everything we know about the brain that macro mental phenomena are all caused by lower-level micro phenomena.

There is nothing mysterious about such bottom-up causation; it is quite common in the physical world. Furthermore, the fact that the mental features are supervenient on neuronal features in no way diminishes their causal efficacy. The solidity of the piston is causally supervenient on its molecular structure, but this does not make solidity epiphenomenal; and similarly, the causal supervenience of my present back pain on micro events in my brain does not make the pain epiphenomenal.

My conclusion is that once you recognize the existence of bottom-up, micro to macro forms of causation, the notion of supervenience no longer does any work in philosophy. The formal features of the relation are already present in the causal sufficiency of the micro-macro forms of causation. And the analogy with ethics is just a source of confusion. The relation of macro mental features of the brain to its micro neuronal features is totally unlike the relation of goodness to good-making features, and it is confusing to lump them together. As Wittgenstein says somewhere, "If you wrap up different kinds of furniture in enough wrapping paper, you can make them all look the same shape."

Chapter 6

The Structure of Consciousness: An Introduction

I have made in passing various claims about the nature of consciousness, and it is now time to attempt a more general account. Such a task can seem both impossibly difficult and ludicrously easy. Difficult because, after all, is not the story of our consciousness the story of our whole life? And easy because, after all, are we not closer to consciousness than to anything else? According to the Cartesian tradition, we have immediate and certain knowledge of our own conscious states, so the job ought to be easy. But it is not. For example, I find it easy to describe the objects on the table in front of me, but how, separately and in addition, would I describe my conscious experience of those objects?

Two subjects are crucial to consciousness, but I will have little to say about them because I do not yet understand them well enough. The first is temporality. Since Kant we have been aware of an asymmetry in the way that consciousness relates to space and to time. Although we experience objects and events as both spatially extended and of temporal duration, our consciousness itself is not experienced as spatial, though it is experienced as temporally extended. Indeed, the spatial metaphors for describing time seem almost inevitable for consciousness as well, as when we speak for example of the "stream of consciousness." Notoriously, phenomenological time does not exactly match real time, but I do not know how to account for the systematic character of the disparities.[1]

The second neglected topic is society. I am convinced that the category of "other people" plays a special role in the *structure* of our conscious experiences, a role unlike that of objects and states of affairs; and I believe that this capacity for assign-