brain and nervous system. What light (we may ask) is thrown by neurophysiology, neurochemistry, and comparative neuroanatomy on such matters as mental illness, learning, three-dimensional vision, and the mental life of dolphins? The answer is, "Considerable light," although neuroscientists will be the first to admit that they have only scratched the surface.

I have included these chapters to provide at least an instructive sampling of the research currently under way in these fields. They are certainly not adequate to introduce an aspiring computer scientist or neuroscientist to these fields. But they will provide some real understanding of how empirical research bears on the philosophical issues discussed in this text. (That is important because, as I hope to make clear, most of those philosophical issues are ultimately empirical in character. They will be decided by the comparative success and the relative progress displayed by alternative scientific research programs.) These chapters will also provide a lasting conceptual framework from which to address future developments concerning the mind. And they may whet your appetite for more empirical information. If they do only that, they will have served their purpose.

The concluding chapter is overtly speculative, as befits a concluding chapter, and opens with an attempt to estimate the distribution of conscious intelligence in the universe at large. Intelligence appears likely to be a fairly widespread phenomenon in the universe, and all advanced instances of it will inevitably face the problem of constructing a useful conception of just what intelligence *is*. That process of self-discovery, to judge from our own case, need not be an easy one. Neither will it be completed in a short period, if indeed it can ever be truly *completed*. But progress is still possible, here, as elsewhere in the human endeavor; and we must be prepared to contemplate revolutions in our conception of what *we* are, just as we have successfully navigated repeated revolutions in our conception of the universe that embeds us. The final section scouts the consequences of such a conceptual revolution for the contents of human self-consciousness.

This concludes my set of promissory notes. Let us now turn to the issues themselves.

Chapter 2

The Ontological Problem (the Mind-Body Problem)

What is the real nature of mental states and processes? In what medium do they take place, and how are they related to the physical world? Will my consciousness survive the disintegration of my physical body? Or will it disappear forever as my brain ceases to function? Is it possible that a purely physical system such as a computer could be constructed so as to enjoy real conscious intelligence? Where do minds come from? What are they?

These are some of the questions we shall confront in this chapter. Which answers we should give to them depends on which theory of mind proves to be the most reasonable theory on the evidence, to have the greatest explanatory power, predictive power, coherence, and simplicity. Let us examine the available theories, and the considerations that weigh for and against each.

1. Dualism

The dualistic approach to mind encompasses several quite different theories, but they are all agreed that the essential nature of conscious intelligence resides in something *nonphysical*, in something forever beyond the scope of sciences like physics, neurophysiology, and computer science. Dualism is not the most widely held view in the current philosophical and scientific community, but it is the most common theory of mind in the public at large, it is deeply entrenched in most of the world's popular religions, and it has been the dominant theory of mind for most of Western history. It is thus an appropriate place to begin our discussion.

Substance Dualism The distinguishing claim of this view is that each mind is a distinct nonphysical thing, an individual 'package' of nonphysical substance, a thing whose identity is independent of any physical body to which it may be temporarily 'attached'. Mental states and activities derive

their special character, on this view, from their being states and activities of this unique, nonphysical substance.

This leaves us wanting to ask for more in the way of a *positive* characterization of the proposed mind-stuff. It is a frequent complaint with the substance dualist's approach that his characterization of it is so far almost entirely negative. This need not be a fatal flaw, however, since we no doubt have much to learn about the underlying nature of mind, and perhaps the deficit here can eventually be made good. On this score, the philosopher René Descartes (1596–1650) has done as much as anyone to provide a positive account of the nature of the proposed mind-stuff, and his views are worthy of examination.

Descartes theorized that reality divides into two basic kinds of substance. The first is ordinary matter, and the essential feature of this kind of substance is that it is extended in space: any instance of it has length, breadth, height, and occupies a determinate position in space. Descartes did not attempt to play down the importance of this type of matter. On the contrary, he was one of the most imaginative physicists of his time, and he was an enthusiastic advocate of what was then called "the mechanical philosophy". But there was one isolated corner of reality he thought could not be accounted for in terms of the mechanics of matter: the conscious reason of Man. This was his motive for proposing a second and radically different kind of substance, a substance that has no spatial extension or spatial position whatever, a substance whose essential feature is the activity of *thinking*. This view is known as *Cartesian dualism*.

As Descartes saw it, the real *you* is not your material body, but rather a nonspatial thinking substance, an individual unit of mind-stuff quite distinct from your material body. This nonphysical mind is in systematic causal interaction with your body. The physical state of your body's sense organs, for example, causes visual/auditory/tactile experiences in your mind. And the desires and decisions of your nonphysical mind cause your body to behave in purposeful ways. Its causal connections to your mind are what make your body yours, and not someone else's.

The main reasons offered in support of this view were straightforward enough. First, Descartes thought that he could determine, by direct introspection alone, that he was essentially a thinking substance and nothing else. And second, he could not imagine how a purely physical system could ever use *language* in a relevant way, or engage in mathematical *reasoning*, as any normal human can. Whether these are good reasons, we shall discuss presently. Let us first notice a difficulty that even Descartes regarded as a problem.

If 'mind-stuff' is so utterly different from 'matter-stuff' in its nature—different to the point that it has no mass whatever, no shape whatever, and no position anywhere in space—then how is it possible for my mind to have any causal influence on my body at all? As Descartes himself was aware (he was one of the first to formulate the law of the conservation of momentum), ordinary matter in space behaves according to rigid laws, and one cannot get bodily movement (= momentum) from nothing. How is this utterly insubstantial 'thinking substance' to have any influence on ponderous matter? How can two such different things be in any sort of causal contact? Descartes proposed a very subtle material substance—'animal spirits'—to convey the mind's influence to the body in general. But this does not provide us with a solution, since it leaves us with the same problem with which we started: how something ponderous and spatial (even 'animal spirits') can interact with something entirely nonspatial.

In any case, the basic principle of division used by Descartes is no longer as plausible as it was in his day. It is now neither useful nor accurate to characterize ordinary matter as that-which-has-extension-in-space. Electrons, for example, are bits of matter, but our best current theories describe the electron as a point-particle with no extension whatever (it even lacks a determinate spatial position). And according to Einstein's theory of gravity, an entire star can achieve this same status, if it undergoes a complete gravitational collapse. If there truly is a division between mind and body, it appears that Descartes did not put his finger on the dividing line.

Such difficulties with Cartesian dualism provide a motive for considering a less radical form of substance dualism, and that is what we find in a view I shall call *popular dualism*. This is the theory that a person is literally a 'ghost in a machine', where the machine is the human body, and the ghost is a spiritual substance, quite unlike physical matter in its internal constitution, but fully possessed of spatial properties even so. In particular, minds are commonly held to be *inside* the bodies they control: inside the head, on most views, in intimate contact with the brain.

This view need not have the difficulties of Descartes'. The mind is right there in contact with the brain, and their interaction can perhaps be understood in terms of their exchanging energy of a form that our science has not yet recognized or understood. Ordinary matter, you may recall, is just a form or manifestation of energy. (You may think of a grain of sand as a great deal of energy condensed or frozen into a small package, according to Einstein's relation, $E = mc^2$.) Perhaps mind-stuff is a well-behaved form or manifestation of energy also, but a different form of it. It is thus *possible* that a dualism of this alternative sort be consistent with familiar laws concerning the conservation of

momentum and energy. This is fortunate for dualism, since those particular laws are very well established indeed.

This view will appeal to many for the further reason that it at least holds out the possibility (though it certainly does not guarantee) that the mind might survive the death of the body. It does not guarantee the mind's survival because it remains possible that the peculiar form of energy here supposed to constitute a mind can be produced and sustained only in conjunction with the highly intricate form of matter we call the brain, and must disintegrate when the brain disintegrates. So the prospects for surviving death are quite unclear even on the assumption that popular dualism is true. But even if survival were a clear consequence of the theory, there is a pitfall to be avoided here. Its promise of survival might be a reason for *wishing* dualism to be true, but it does not constitute a reason for *believing* that it *is* true. For that, we would need independent empirical evidence that minds do indeed survive the permanent death of the body. Regrettably, and despite the exploitative blatherings of the supermarket tabloids (**TOP DOCS PROVE LIFE AFTER DEATH!!!**), we possess no such evidence.

As we shall see later in this section, when we turn to evaluation, positive evidence for the existence of this novel, nonmaterial, thinking *substance* is in general on the slim side. This has moved many dualists to articulate still less extreme forms of dualism, in hopes of narrowing further the gap between theory and available evidence.

Property Dualism The basic idea of the theories under this heading is that while there is no *substance* to be dealt with here beyond the physical brain, the brain has a special set of *properties* possessed by no other kind of physical object. It is these special properties that are nonphysical: hence the term *property dualism*. The properties in question are the ones you would expect: the property of having a pain, of having a sensation of red, of thinking that *P*, of desiring that *Q*, and so forth. These are the properties that are characteristic of conscious intelligence. They are held to be nonphysical in the sense that they cannot ever be reduced to or explained solely in terms of the concepts of the familiar physical sciences. They will require a wholly new and autonomous science—the 'science of mental phenomena'—if they are ever to be adequately understood.

From here, important differences among the positions emerge. Let us begin with what is perhaps the oldest version of property dualism: *epiphenomenalism*. This term is rather a mouthful, but its meaning is simple. The Greek prefix "epi-" means "above", and the position at issue holds that mental phenomena are not a part of the physical

phenomena in the brain that ultimately determine our actions and behavior, but rather ride 'above the fray'. Mental phenomena are thus *epi*phenomena. They are held to just appear or emerge when the growing brain passes a certain level of complexity.

But there is more. The epiphenomenalist holds that while mental phenomena are caused to occur by the various activities of the brain, *they do not have any causal effects in turn*. They are entirely impotent with respect to causal effects on the physical world. They are *mere* epiphenomena. (To fix our ideas, a vague metaphor may be helpful here. Think of our conscious mental states as little sparkles of shimmering light that occur on the wrinkled surface of the brain, sparkles which are caused to occur by physical activity in the brain, but which have no causal effects on the brain in return.) This means that the universal conviction that one's actions are determined by one's desires, decisions, and volitions is false! One's actions are exhaustively determined by physical events in the brain, which events *also* cause the epiphenomena we call desires, decisions, and volitions. There is therefore a constant conjunction between volitions and actions. But according to the epiphenomenalist, it is mere illusion that the former cause the latter.

What could motivate such a strange view? In fact, it is not too difficult to understand why someone might take it seriously. Put yourself in the shoes of a neuroscientist who is concerned to trace the origins of behavior back up the motor nerves to the active cells in the motor cortex of the cerebrum, and to trace in turn their activity into inputs from other parts of the brain, and from the various sensory nerves. She finds a thoroughly physical system of awesome structure and delicacy, and much intricate activity, all of it unambiguously chemical or electrical in nature, and she finds no hint at all of any nonphysical inputs of the kind that substance dualism proposes. What is she to think? From the standpoint of her researches, human behavior is exhaustively a function of the activity of the physical brain. And this opinion is further supported by her confidence that the brain has the behavior-controlling features it does exactly because those features have been ruthlessly selected for during the brain's long evolutionary history. In sum, the seat of human behavior appears entirely physical in its constitution, in its origins, and in its internal activities.

On the other hand, our neuroscientist has the testimony of her own introspection to account for as well. She can hardly deny that she has experiences, beliefs, and desires, nor that they are connected in some way with her behavior. One bargain that can be struck here is to admit the *reality* of mental properties, as nonphysical properties, but demote them to the status of impotent epiphenomena that have nothing to do

with the scientific explanation of human and animal behavior. This is the position the epiphenomenalist takes, and the reader can now perceive the rationale behind it. It is a bargain struck between the desire to respect a rigorously scientific approach to the explanation of behavior, and the desire to respect the testimony of introspection.

The epiphenomenalist's 'demotion' of mental properties—to causally impotent by-products of brain activity—has seemed too extreme for most property dualists, and a theory closer to the convictions of common sense has enjoyed somewhat greater popularity. This view, which we may call *interactionist property dualism*, differs from the previous view in only one essential respect: the interactionist asserts that mental properties do indeed have causal effects on the brain, and thereby, on behavior. The mental properties of the brain are an integrated part of the general causal fray, in systematic interaction with the brain's physical properties. One's actions, therefore, are held to be caused by one's desires and volitions after all.

As before, mental properties are here said to be *emergent* properties, properties that do not appear at all until ordinary physical matter has managed to organize itself, through the evolutionary process, into a system of sufficient complexity. Examples of properties that are emergent in this sense would be the property of being *solid*, the property of being *colored*, and the property of being *alive*. All of these require matter to be suitably organized before they can be displayed. With this much, any materialist will agree. But any property dualist makes the further claim that mental states and properties are *irreducible*, in the sense that they are not just organizational features of physical matter, as are the examples cited. They are said to be novel properties beyond prediction or explanation by physical science.

This last condition—the irreducibility of mental properties—is an important one, since this is what makes the position a dualist position. But it sits poorly with the joint claim that mental properties emerge from nothing more than the organizational achievements of physical matter. If that is how mental properties are produced, then one would expect a physical account of them to be possible. The simultaneous claim of evolutionary emergence *and* physical irreducibility is prima facie puzzling.

A property dualist is not absolutely bound to insist on both claims. He could let go the thesis of evolutionary emergence, and claim that mental properties are *fundamental* properties of reality, properties that have been here from the universe's inception, properties on a par with length, mass, electric charge, and other fundamental properties. There is even an historical precedent for a position of this kind. At the turn of this century it was still widely believed that electromagnetic phenomena (such as electric charge and magnetic attraction) were just an unusually subtle manifestation of purely *mechanical* phenomena. Some scientists thought that a reduction of electromagnetics to mechanics was more or less in the bag. They thought that radio waves, for example, would turn out to be just travelling oscillations in a very subtle but jellylike aether that fills space everywhere. But the aether turned out not to exist. So electromagnetic properties turned out to be fundamental properties in their own right, and we were forced to add electric charge to the existing list of fundamental properties (mass, length, and duration).

Perhaps mental properties enjoy a status like that of electromagnetic properties: irreducible, but not emergent. Such a view may be called *elemental-property dualism*, and it has the advantage of clarity over the previous view. Unfortunately, the parallel with electromagnetic phenomena has one very obvious failure. Unlike electromagnetic properties, which are displayed at all levels of reality from the subatomic level on up, mental properties are displayed only in large physical systems that have evolved a very complex internal organization. The case for the evolutionary emergence of mental properties through the organization of matter is extremely strong. They do not appear to be basic or elemental at all. This returns us, therefore, to the issue of their irreducibility. Why should we accept this most basic of the dualist's claims? Why be a dualist?

Arguments for Dualism Here we shall examine some of the main considerations commonly offered in support of dualism. Criticism will be postponed for a moment so that we may appreciate the collective force of these supporting considerations.

A major source of dualistic convictions is the religious belief many of us bring to these issues. Each of the major religions is in its way a theory about the cause or purpose of the universe, and Man's place within it, and many of them are committed to the notion of an immortal soul—that is, to some form of substance dualism. Supposing that one is consistent, to consider disbelieving dualism is to consider disbelieving one's religious heritage, and some of us find that difficult to do. Call this the *argument from religion*.

A more universal consideration is the *argument from introspection*. The fact is, when you center your attention on the contents of your consciousness, you do not clearly apprehend a neural network pulsing with electrochemical activity: you apprehend a flux of thoughts, sensations, desires, and emotions. It seems that mental states and properties, as revealed in introspection, could hardly be more different from physical

states and properties if they tried. The verdict of introspection, therefore, seems strongly on the side of some form of dualism—on the side of property dualism, at a minimum.

A cluster of important considerations can be collected under the *argument from irreducibility*. Here one points to a variety of mental phenomena where it seems clear that no purely physical explanation could possibly account for what is going on. Descartes has already cited our ability to use language in a way that is relevant to our changing circumstances, and he was impressed also with our faculty of Reason, particularly as it is displayed in our capacity for mathematical reasoning. These abilities, he thought, must surely be beyond the capacity of any physical system. More recently, the introspectible qualities of our sensations (sensory 'qualia'), and the meaningful content of our thoughts and beliefs, have also been cited as phenomena that will forever resist reduction to the physical. Consider, for example, seeing the color or smelling the fragrance of a rose. A physicist or chemist might know everything about the molecular structure of the rose, and of the human brain, argues the dualist, but that knowledge would not enable him to predict or anticipate the quality of these inexpressible experiences.

Finally, paraphsychological phenomena are occasionally cited in favor of dualism. Telepathy (mind reading), precognition (seeing the future), telekinesis (thought control of material objects), and clairvoyance (knowledge of distant objects) are all awkward to explain within the normal confines of psychology and physics. If these phenomena are real, they might well be reflecting the superphysical nature that the dualist ascribes to the mind. Trivially they are *mental* phenomena, and if they are also forever beyond physical explanation, then at least some mental phenomena must be irreducibly nonphysical.

Collectively, these considerations may seem compelling. But there are serious criticisms of each, and we must examine them as well. Consider first the argument from religion. There is certainly nothing wrong in principle with appealing to a more general theory that bears on the case at issue, which is what the appeal to religion amounts to. But the appeal can only be as good as the scientific credentials of the religion(s) being appealed to, and here the appeals tend to fall down rather badly. In general, attempts to decide scientific questions by appeal to religious orthodoxy have a very sorry history. That the stars are other suns, that the earth is not the center of the universe, that diseases are caused by microorganisms, that the earth is billions of years old, that life is a physicochemical phenomenon; all of these crucial insights were strongly and sometimes viciously resisted, because the dominant religion of the time happened to think otherwise. Giordano Bruno was

burned at the stake for urging the first view; Galileo was forced by threat of torture in the Vatican's basement to recant the second view; the firm belief that disease was a punishment visited by the Devil allowed public health practices that brought chronic plagues to most of the cities of Europe; and the age of the earth and the evolution of life were forced to fight an uphill battle against religious prejudice even in an age of supposed enlightenment.

History aside, the almost universal opinion that one's own religious convictions are the reasoned outcome of a dispassionate evaluation of all of the major alternatives is almost demonstrably false for humanity in general. If that really were the genesis of most people's convictions, then one would expect the major faiths to be distributed more or less randomly or evenly over the globe. But in fact they show a very strong tendency to cluster: Christianity is centered in Europe and the Americas, Islam in Africa and the Middle East, Hinduism in India, and Buddhism in the Orient. Which illustrates what we all suspected anyway: that *social forces* are the primary determinants of religious belief for people in general. To decide scientific questions by appeal to religious orthodoxy would therefore be to put social forces in place of empirical evidence. For all of these reasons, professional scientists and philosophers concerned with the nature of mind generally do their best to keep religious appeals out of the discussion entirely.

The argument from introspection is a much more interesting argument, since it tries to appeal to the direct experience of everyman. But the argument is deeply suspect, in that it assumes that our faculty of inner observation or introspection reveals things as they really are in their innermost nature. This assumption is suspect because we already know that our other forms of observation—sight, hearing, touch, and so on—do no such thing. The red surface of an apple does not *look* like a matrix of molecules reflecting photons at certain critical wavelengths, but that is what it is. The sound of a flute does not *sound* like a sinusoidal compression wave train in the atmosphere, but that is what it is. The warmth of the summer air does not *feel* like the mean kinetic energy of millions of tiny molecules, but that is what it is. If one's pains and hopes and beliefs do not *introspectively* seem like electrochemical states in a neural network, that may be only because our faculty of introspection, like our other senses, is not sufficiently penetrating to reveal such hidden details. Which is just what one would expect anyway. The argument from introspection is therefore entirely without force, unless we can somehow argue that the faculty of introspection is quite different from all other forms of observation.

The argument from irreducibility presents a more serious challenge, but here also its force is less than first impression suggests. Consider first our capacity for mathematical reasoning which so impressed Des-

cartes. The last ten years have made available, to anyone with fifty dollars to spend, electronic calculators whose capacity for mathematical reasoning—the calculational part, at least—far surpasses that of any normal human. The fact is, in the centuries since Descartes' writings, philosophers, logicians, mathematicians, and computer scientists have managed to isolate the general principles of mathematical reasoning, and electronics engineers have created machines that compute in accord with those principles. The result is a hand-held object that would have astonished Descartes. This outcome is impressive not just because machines have proved capable of some of the capacities boasted by human reason, but because some of those achievements invade areas of human reason that past dualistic philosophers have held up as forever closed to mere physical devices.

Although debate on the matter remains open, Descartes' argument from language use is equally dubious. The notion of a *computer language* is by now a commonplace: consider BASIC, PASCAL, FORTRAN, APL, LISP, and so on. Granted, these artificial 'languages' are much simpler in structure and content than human natural language, but the differences may be differences only of degree, and not of kind. As well, the theoretical work of Noam Chomsky and the generative grammar approach to linguistics have done a great deal to explain the human capacity for language use in terms that invite simulation by computers. I do not mean to suggest that truly conversational computers are just around the corner. We have a great deal yet to learn, and fundamental problems yet to solve (mostly having to do with our capacity for inductive or theoretical reasoning). But recent progress here does nothing to support the claim that language use must be forever impossible for a purely physical system. On the contrary, such a claim now appears rather arbitrary and dogmatic, as we shall see in chapter 6.

The next issue is also a live problem: How can we possibly hope to explain or to predict the intrinsic qualities of our sensations, or the meaningful content of our beliefs and desires, in purely physical terms? This is a major challenge to the materialist. But as we shall see in later sections, active research programs are already under way on both problems, and positive suggestions are being explored. It is in fact not impossible to imagine how such explanations might go, though the materialist cannot yet pretend to have solved either problem. Until he does, the dualist will retain a bargaining chip here, but that is about all. What the dualists need in order to establish their case is the conclusion that a physical reduction is outright impossible, and that is a conclusion they have failed to establish. Rhetorical questions, like the one that opens this paragraph, do not constitute arguments. And it is equally difficult, note, to imagine how the relevant phenomena could

be explained or predicted solely in terms of the substance dualist's nonphysical mind-stuff. The explanatory problem here is a major challenge to everybody, not just to the materialist. On this issue then, we have a rough standoff.

The final argument in support of dualism urged the existence of parapsychological phenomena such as telepathy and telekinesis, the point being that such mental phenomena are (a) real, and (b) beyond purely physical explanation. This argument is really another instance of the argument from irreducibility discussed above, and as before, it is not entirely clear that such phenomena, even if real, must forever escape a purely physical explanation. The materialist can already suggest a possible mechanism for telepathy, for example. On his view, thinking is an electrical activity within the brain. But according to electromagnetic theory, such changing motions of electric charges must produce electromagnetic waves radiating at the speed of light in all directions, waves that will contain information about the electrical activity that produced them. Such waves can subsequently have effects on the electrical activity of other brains, that is, on their thinking. Call this the 'radio transmitter/receiver' theory of telepathy.

I do not for a moment suggest that this theory is true: the electromagnetic waves emitted by the brain are fantastically weak (billions of times weaker than the ever present background electromagnetic flux produced by commercial radio stations), and they are almost certain to be hopelessly jumbled together as well. This is one reason why, in the absence of systematic, compelling, and repeatable evidence for the existence of telepathy, one must doubt its possibility. But it is significant that the materialist has the theoretical resources to suggest a detailed possible explanation of telepathy, if it were real, which is more than any dualist has so far done. It is not at all clear, then, that the materialist *must* be at an explanatory disadvantage in these matters. Quite the reverse.

Put the preceding aside, if you wish, for the main difficulty with the argument from parapsychological phenomena is much, much simpler. Despite the endless pronouncements and anecdotes in the popular press, and despite a steady trickle of serious research on such things, there is no significant or trustworthy evidence that such phenomena even exist. The wide gap between popular conviction on this matter, and the actual evidence, is something that itself calls for research. For there is not a single parapsychological effect that can be repeatedly or reliably produced in any laboratory suitably equipped to perform and control the experiment. Not one. Honest researchers have been repeatedly hoodwinked by 'psychic' charlatans with skills derived from the magician's trade, and the history of the subject is largely a history

of gullibility, selection of evidence, poor experimental controls, and outright fraud by the occasional researcher as well. If someone really does discover a repeatable parapsychological effect, then we shall have to reevaluate the situation, but as things stand, there is nothing here to support a dualist theory of mind.

Upon critical examination, the arguments in support of dualism lose much of their force. But we are not yet done: there are arguments against dualism, and these also require examination.

Arguments against Dualism The first argument against dualism urged by the materialists appeals to the greater *simplicity* of their view. It is a principle of rational methodology that, if all else is equal, the simpler of two competing hypotheses should be preferred. This principle is sometimes called "Ockham's Razor"—after William of Ockham, the medieval philosopher who first enunciated it—and it can also be expressed as follows: "Do not multiply entities beyond what is strictly necessary to explain the phenomena." The materialist postulates only one kind of substance (physical matter), and one class of properties (physical properties), whereas the dualist postulates two kinds of matter and/or two classes of properties. And to no explanatory advantage, charges the materialist.

This is not yet a decisive point against dualism, since neither dualism nor materialism can yet explain all of the phenomena to be explained. But the objection does have some force, especially since there is no doubt at all that physical matter exists, while spiritual matter remains a tenuous hypothesis.

If this latter hypothesis brought us some definite explanatory advantage obtainable in no other way, then we would happily violate the demand for simplicity, and we would be right to do so. But it does not, claims the materialist. In fact, the advantage is just the other way around, he argues, and this brings us to the second objection to dualism: the relative *explanatory impotence* of dualism as compared to materialism.

Consider, very briefly, the explanatory resources already available to the neurosciences. We know that the brain exists and what it is made of. We know much of its microstructure: how the neurons are organized into systems and how distinct systems are connected to one another, to the motor nerves going out to the muscles, and to the sensory nerves coming in from the sense organs. We know much of their microchemistry: how the nerve cells fire tiny electrochemical pulses along their various fibers, and how they make other cells fire also, or cease firing. We know some of how such activity processes sensory information, selecting salient or subtle bits to be sent on to higher systems. And we know some of how such activity initiates and coordinates bodily behavior. Thanks mainly to neurology (the branch of medicine concerned with brain pathology), we know a great deal about the correlations between damage to various parts of the human brain, and various behavioral and cognitive deficits from which the victims suffer. There are a great many isolated deficits—some gross, some subtle—that are familiar to neurologists (inability to speak, or to read, or to understand speech, or to recognize faces, or to add/subtract, or to move a certain limb, or to put information into long-term memory, and so on), and their appearance is closely tied to the occurrence of damage to very specific parts of the brain.

Nor are we limited to cataloguing traumas. The growth and development of the brain's microstructure is also something that neuroscience has explored, and such development appears to be the basis of various kinds of learning by the organism. Learning, that is, involves lasting chemical and physical changes in the brain. In sum, the neuroscientist can tell us a great deal about the brain, about its constitution and the physical laws that govern it; he can already explain much of our behavior in terms of the physical, chemical, and electrical properties of the brain; and he has the theoretical resources available to explain a good deal more as our explorations continue. (We shall take a closer look at neurophysiology and neuropsychology in chapter 7.)

Compare now what the neuroscientist can tell us about the brain, and what he can do with that knowledge, with what the dualist can tell us about spiritual substance, and what he can do with those assumptions. Can the dualist tell us anything about the internal constitution of mind-stuff? Of the nonmaterial elements that make it up? Of the laws that govern their behavior? Of the mind's structural connections with the body? Of the manner of its operations? Can he explain human capacities and pathologies in terms of its structures and its defects? The fact is, the dualist can do none of these things, because no detailed theory of mind-stuff has ever been formulated. Compared to the rich resources and explanatory successes of current materialism, dualism is less a theory of mind than it is an empty space waiting for a genuine theory of mind to be put in it.

Thus argues the materialist. But again, this is not a completely decisive point against dualism. The dualist can admit that the brain plays a major role in the administration of both perception and behavior—on his view the brain is the *mediator* between the mind and the body—but he may attempt to argue that the materialist's current successes and future explanatory prospects concern only the mediative functions of the brain, not the *central* capacities of the nonphysical mind, capacities such as reason, emotion, and consciousness itself. On these latter topics, he may argue, both dualism *and* materialism currently draw a blank.

But this reply is not a very good one. So far as the capacity for reasoning is concerned, machines already exist that execute in minutes sophisticated deductive and mathematical calculations that would take a human a lifetime to execute. And so far as the other two mental capacities are concerned, studies of such things as depression, motivation, attention, and sleep have revealed many interesting and puzzling facts about the neurochemical and neurodynamical basis of both emotion and consciousness. The *central* capacities, no less than the peripheral, have been addressed with profit by various materialist research programs.

In any case, the (substance) dualist's attempt to draw a sharp distinction between the unique 'mental' capacities proper to the nonmaterial mind, and the merely mediative capacities of the brain, prompts an argument that comes close to being an outright refutation of (substance) dualism. If there really is a distinct entity in which reasoning, emotion, and consciousness take place, and if that entity is dependent on the brain for nothing more than sensory experiences as input and volitional executions as output, *then one would expect reason, emotion, and consciousness to be relatively invulnerable to direct control or pathology by manipulation or damage to the brain.* But in fact the exact opposite is true. Alcohol, narcotics, or senile degeneration of nerve tissue will impair, cripple, or even destroy one's capacity for rational thought. Psychiatry knows of hundreds of emotion-controlling chemicals (lithium, chlorpromazine, amphetamine, cocaine, and so on) that do their work when vectored into the brain. And the vulnerability of consciousness to the anesthetics, to caffeine, and to something as simple as a sharp blow to the head, shows its very close dependence on neural activity in the brain. All of this makes perfect sense if reason, emotion, and consciousness are activities of the brain itself. But it makes very little sense if they are activities of something else entirely.

We may call this the argument from the *neural dependence* of all known mental phenomena. Property dualism, note, is not threatened by this argument, since, like materialism, property dualism reckons the brain as the seat of all mental activity. We shall conclude this section, however, with an argument that cuts against both varieties of dualism: the argument from *evolutionary history.*

What is the origin of a complex and sophisticated species such as ours? What, for that matter, is the origin of the dolphin, the mouse, or the housefly? Thanks to the fossil record, comparative anatomy, and the biochemistry of proteins and nucleic acids, there is no longer any significant doubt on this matter. Each existing species is a surviving type from a number of variations on an earlier type of organism; each earlier type is in turn a surviving type from a number of variations on a still earlier type of organism; and so on down the branches of the evolutionary tree until, some three billion years ago, we find a trunk of just one or a handful of very simple organisms. These organisms, like their more complex offspring, are just self-repairing, self-replicating, energy-driven molecular structures. (That evolutionary trunk has its own roots in an earlier era of purely chemical evolution, in which the molecular elements of life were themselves pieced together.) The mechanism of development that has structured this tree has two main elements: (1) the occasional blind variation in types of reproducing creature, and (2) the selective survival of some of these types due to the relative reproductive advantage enjoyed by individuals of those types. Over periods of geological time, such a process can produce an enormous variety of organisms, some of them very complex indeed.

For purposes of our discussion, the important point about the standard evolutionary story is that the human species and all of its features are the wholly physical outcome of a purely physical process. Like all but the simplest of organisms, we have a nervous system. And for the same reason: a nervous system permits the discriminative guidance of behavior. But a nervous system is just an active matrix of cells, and a cell is just an active matrix of molecules. We are notable only in that our nervous system is more complex and powerful than those of our fellow creatures. Our inner nature differs from that of simpler creatures in degree, but not in kind.

If this is the correct account of our origins, then there seems neither need, nor room, to fit any nonphysical substances or properties into our theoretical account of ourselves. We are creatures of matter. And we should learn to live with that fact.

Arguments like these have moved most (but not all) of the professional community to embrace some form of materialism. This has not produced much unanimity, however, since the differences between the several materialist positions are even wider than the differences that divide dualism. The next four sections explore these more recent positions.

Suggested Readings

On Substance Dualism

Descartes, René, *The Meditations*, meditation II.
Descartes, René, *Discourse on Method*, part 5.
Eccles, Sir John C., *The Self and Its Brain*, with Sir Karl Popper (New York: Springer-Verlag, 1977).

On Property Dualism

Popper, Sir Karl, *The Self and Its Brain*, with Sir John C. Eccles (New York: Springer-Verlag, 1977).

Margolis, Joseph, *Persons and Minds: The Prospects of Nonreductive Materialism* (Dordrecht-Holland: Reidel, 1978).

Jackson, Frank, "Epiphenomenal Qualia," *The Philosophical Quarterly*, vol. 32, no. 127 (April, 1982).

Nagel, Thomas, "What Is It Like to Be a Bat?" *Philosophical Review*, vol. LXXXIII (1974). Reprinted in *Readings in Philosophy of Psychology*, vol. I, ed. N. Block (Cambridge, MA: Harvard University Press, 1980).

2. *Philosophical Behaviorism*

Philosophical behaviorism reached the peak of its influence during the first and second decades after World War II. It was jointly motivated by at least three intellectual fashions. The first motivation was a reaction against dualism. The second motivation was the Logical Positivists' idea that the meaning of any sentence was ultimately a matter of the observable circumstances that would tend to verify or confirm that sentence. And the third motivation was a general assumption that most, if not all, philosophical problems are the result of linguistic or conceptual confusion, and are to be solved (or dissolved) by careful analysis of the language in which the problem is expressed.

In fact, philosophical behaviorism is not so much a theory about what mental states are (in their inner nature) as it is a theory about how to analyze or to understand the vocabulary we use to talk about them. Specifically, the claim is that talk about emotions and sensations and beliefs and desires is not talk about ghostly inner episodes, but is rather a shorthand way of talking about actual and potential patterns of *behavior*. In its strongest and most straightforward form, philosophical behaviorism claims that any sentence about a mental state can be paraphrased, without loss of meaning, into a long and complex sentence about what observable behavior *would* result if the person in question were in this, that, or the other observable circumstance.

A helpful analogy here is the dispositional property, *being soluble*. To say that a sugar cube is soluble is not to say that the sugar cube enjoys some ghostly inner state. It is just to say that *if* the sugar cube were put in water, then it *would* dissolve. More strictly,

"x is water soluble"

is equivalent by definition to

"if x were put in unsaturated water, x would dissolve."

This is one example of what is called an "operational definition". The term "soluble" is defined in terms of certain operations or tests that would reveal whether or not the term actually applies in the case to be tested.

According to the behaviorist, a similar analysis holds for mental states such as "wants a Caribbean holiday", save that the analysis is much richer. To say that Anne wants a Caribbean holiday is to say that (1) if asked whether that is what she wants, she would answer yes, and (2) if given new holiday brochures for Jamaica and Japan, she would peruse the ones for Jamaica first, and (3) if given a ticket on this Friday's flight to Jamaica, she would go, and so on and so on.

Unlike solubility, claims the behaviorist, most mental states are *multi-tracked* dispositions. But dispositions they remain.

There is therefore no point in worrying about the 'relation' between the mind and the body, on this view. To talk about Marie Curie's mind, for example, is not to talk about some 'thing' that she 'possesses'; it is to talk about certain of her extraordinary capacities and dispositions. The mind-body problem, concludes the behaviorist, is a pseudoproblem.

Behaviorism is clearly consistent with a materialist conception of human beings. Material objects can have dispositional properties, even multitracked ones, so there is no necessity to embrace dualism to make sense of our psychological vocabulary. (It should be pointed out, however, that behaviorism is strictly consistent with dualism also. Even if philosophical behaviorism were true, it would remain possible that our multitracked dispositions are grounded in immaterial mind-stuff rather than in molecular structures. This is not a possibility that most behaviorists took seriously, however, for the many reasons outlined at the end of the preceding section.)

Philosophical behaviorism, unfortunately, had two major flaws that made it awkward to believe, even for its defenders. It evidently ignored, and even denied, the 'inner' aspect of our mental states. To have a pain, for example, seems to be not merely a matter of being inclined to moan, to wince, to take aspirin, and so on. Pains also have an intrinsic qualitative nature (a horrible one) that is revealed in introspection, and any theory of mind that ignores or denies such *qualia* is simply derelict in its duty.

This problem received much attention from behaviorists, and serious attempts were made to solve it. The details take us deeply into semantical problems, however, so we shall postpone further discussion of this difficulty until chapter 3.

The second flaw emerged when behaviorists attempted to specify in detail the multitracked disposition said to constitute any given mental state. The list of conditionals necessary for an adequate analysis of "wants a Caribbean holiday", for example, seemed not just to be long, but to be indefinitely or even infinitely long, with no finite way of specifying the elements to be included. And no term can be well-defined whose *definiens* is open-ended and unspecific in this way. Further, each conditional of the long analysis was suspect on its own. Supposing that Anne does want a Caribbean holiday, conditional (1) above will be true only if she isn't *secretive* about her holiday fantasies; conditional (2) will be true only if she isn't already *bored* with the Jamaica brochures; conditional (3) will be true only if she doesn't *believe* the Friday flight will be hijacked, and so forth. But to repair each conditional by adding in the relevant qualification would be to reintroduce a series of *mental* elements into the business end of the definition, and we would no longer be defining the mental solely in terms of publicly observable circumstances and behavior.

So long as behaviorism seemed the only alternative to dualism, philosophers were prepared to struggle with these flaws in hopes of repairing or defusing them. However, three more materialist theories rose to prominence during the late fifties and sixties, and the flight from behaviorism was swift.

(I close this section with a cautionary note. The *philosophical* behaviorism discussed above is to be sharply distinguished from the *methodological* behaviorism that has enjoyed such a wide influence within psychology. In its bluntest form, this latter view urges that any new theoretical terms invented by the science of psychology *should be* operationally defined, in order to guarantee that psychology maintains a firm contact with empirical reality. Philosophical behaviorism, by contrast, claims that all of the common-sense psychological terms in our prescientific vocabulary *already* get whatever meaning they have from (tacit) operational definitions. The two views are logically distinct, and the methodology might be a wise one, for new theoretical terms, even though the correlative analysis of common-sense mental terms is wrong.)

Suggested Readings

Ryle, Gilbert, *The Concept of Mind* (London: Hutchinson & Company, 1949), chapters I and V.

Malcolm, Norman, "Wittgenstein's *Philosophical Investigations*," *Philosophical Review*, vol. XLVII (1956). Reprinted in *The Philosophy of Mind*, ed. V. C. Chappell (Englewood Cliffs, NJ: Prentice-Hall, 1962).

3. *Reductive Materialism (the Identity Theory)*

Reductive materialism, more commonly known as *the identity theory,* is the most straightforward of the several materialist theories of mind. Its central claim is simplicity itself: Mental states *are* physical states of the brain. That is, each type of mental state or process is *numerically identical with* (is one and the very same thing as) some type of physical state or process within the brain or central nervous system. At present we do not know enough about the intricate functionings of the brain actually to state the relevant identities, but the identity theory is committed to the idea that brain research will eventually reveal them. (Partly to help us evaluate that claim, we shall examine current brain research in chapter 7.)

Historical Parallels As the identity theorist sees it, the result here predicted has familiar parallels elsewhere in our scientific history. Consider sound. We now know that sound is just a train of compression waves traveling through the air, and that the property of being high pitched is identical with the property of having a high oscillatory frequency. We have learned that light is just electromagnetic waves, and our best current theory says that the color of an object is identical with a triplet of reflectance efficiencies the object has, rather like a musical chord that it strikes, though the 'notes' are struck in electromagnetic waves instead of in sound waves. We now appreciate that the warmth or coolness of a body is just the energy of motion of the molecules that make it up: warmth is identical with high average molecular kinetic energy, and coolness is identical with low average molecular kinetic energy. We know that lightning is identical with a sudden large-scale discharge of electrons between clouds, or between the atmosphere and the ground. What we now think of as 'mental states,' argues the identity theorist, are identical with brain states in exactly the same way.

Intertheoretic Reduction These illustrative parallels are all cases of successful *intertheoretic reduction.* That is, they are all cases where a new and very powerful theory turns out to entail a set of propositions and principles that mirror perfectly (or almost perfectly) the propositions and principles of some older theory or conceptual framework. The relevant principles entailed by the new theory have the same structure as the corresponding principles of the old framework, and they apply in exactly the same cases. The only difference is that where the old principles contained (for example) the notions of "heat", "is hot", and "is cold", the new principles contain instead the notions of "total molecular kinetic energy", "has a high mean molecular kinetic energy", and "has a low mean molecular kinetic energy".

If the new framework is far better than the old at explaining and predicting phenomena, then we have excellent reason for believing that the theoretical terms of the *new* framework are the terms that describe reality correctly. But if the old framework worked adequately, so far as it went, and if it parallels a portion of the new theory in the systematic way described, then we may properly conclude that the old terms and the new terms refer to the very same things, or express the very same properties. We conclude that we have apprehended the very same reality that is incompletely described by the old framework, but with a new and more penetrating conceptual framework. And we announce what philosophers of science call "intertheoretic identities": light *is* electromagnetic waves, temperature *is* mean molecular kinetic energy, and so forth.

The examples of the preceding two paragraphs share one more important feature in common. They are all cases where the things or properties on the receiving end of the reduction are *observable* things and properties within our *common-sense* conceptual framework. They show that intertheoretic reduction occurs not only between conceptual frameworks in the theoretical stratosphere: common-sense observables can also be reduced. There would therefore be nothing particularly surprising about a reduction of our familiar introspectible mental states to physical states of the brain. All that would be required would be that an explanatorily successful neuroscience develop to the point where it entails a suitable 'mirror image' of the assumptions and principles that constitute our common-sense conceptual framework for mental states, an image where brain-state terms occupy the positions held by mental-state terms in the assumptions and principles of common sense. If this (rather demanding) condition were indeed met, then, as in the historical cases cited, we would be justified in announcing a reduction, and in asserting the identity of mental states with brain states.

Arguments for the Identity Theory What reasons does the identity theorist have for believing that neuroscience will eventually achieve the strong conditions necessary for the reduction of our 'folk' psychology? There are at least four reasons, all directed at the conclusion that the correct account of human-behavior-and-its-causes must reside in the physical neurosciences.

We can point first to the purely physical origins and ostensibly physical constitution of each individual human. One begins as a genetically

programmed monocellular organization of molecules (a fertilized ovum), and one develops from there by the accretion of further molecules whose structure and integration is controlled by the information coded in the DNA molecules of the cell nucleus. The result of such a process would be a purely physical system whose behavior arises from its internal operations and its interactions with the rest of the physical world. And those behavior-controlling internal operations are precisely what the neurosciences are about.

This argument coheres with a second argument. The origins of each *type* of animal also appear exhaustively physical in nature. The argument from evolutionary history discussed earlier (p. 20) lends further support to the identity theorist's claim, since evolutionary theory provides the only serious explanation we have for the behavior-controlling capacities of the brain and central nervous system. Those systems were selected for because of the many advantages (ultimately, the reproductive advantage) held by creatures whose behavior was thus controlled. Again our behavior appears to have its basic causes in neural activity.

The identity theorist finds further support in the argument, discussed earlier, from the neural dependence of all known mental phenomena (see p. 20). This is precisely what one should expect, if the identity theory is true. Of course, systematic neural dependence is also a consequence of property dualism, but here the identity theorist will appeal to considerations of simplicity. Why admit two radically different classes of properties and operations if the explanatory job can be done by one?

A final argument derives from the growing success of the neurosciences in unraveling the nervous systems of many creatures and in explaining their behavioral capacities and deficits in terms of the structures discovered. The preceding arguments all suggest that neuroscience should be successful in this endeavor, and the fact is that the continuing history of neuroscience bears them out. Especially in the case of very simple creatures (as one would expect), progress has been rapid. And progress has also been made with humans, though for obvious moral reasons exploration must be more cautious and circumspect. In sum, the neurosciences have a long way to go, but progress to date provides substantial encouragement to the identity theorist.

Even so, these arguments are far from decisive in favor of the identity theory. No doubt they do provide an overwhelming case for the idea that the causes of human and animal behavior are essentially physical in nature, but the identity theory claims more than just this. It claims that neuroscience will discover a taxonomy of neural states that stand in a one-to-one correspondence with the mental states of our commonsense taxonomy. Claims for intertheoretic identity will be justified only if such a match-up can be found. But nothing in the preceding arguments

guarantees that the old and new frameworks will match up in this way, even if the new framework is a roaring success at explaining and predicting our behavior. Furthermore, there are arguments from other positions within the materialist camp to the effect that such convenient match-ups are rather unlikely. Before exploring those, however, let us look at some more traditional objections to the identity theory.

Arguments against the Identity Theory

We may begin with the argument from introspection discussed earlier. Introspection reveals a domain of thoughts, sensations, and emotions, not a domain of electrochemical impulses in a neural network. Mental states and properties, as revealed in introspection, appear radically different from any neurophysiological states and properties. How could they possibly be the very same things?

The answer, as we have already seen, is, "Easily." In discriminating red from blue, sweet from sour, and hot from cold, our external senses are actually discriminating between subtle differences in intricate electromagnetic, stereochemical, and micromechanical properties of physical objects. But our senses are not sufficiently penetrating to reveal on their own the detailed nature of those intricate properties. That requires theoretical research and experimental exploration with specially designed instruments. The same is presumably true of our 'inner' sense: introspection. It may discriminate efficiently between a great variety of neural states, without being able to reveal on its own the detailed nature of the states being discriminated. Indeed, it would be faintly miraculous if it did reveal them, just as miraculous as if unaided sight were to reveal the existence of interacting electric and magnetic fields whizzing by with an oscillatory frequency of a million billion hertz and a wavelength of less than a millionth of a meter. For despite 'appearances', that is what light is. The argument from introspection, therefore, is quite without force.

The next objection argues that the identification of mental states with brain states would commit us to statements that are literally unintelligible, to what philosophers have called "category errors", and that the identification is therefore a case of sheer conceptual confusion. We may begin the discussion by noting a most important law concerning numerical identity. Leibniz' Law states that two items are numerically identical just in case any property had by either one of them is also had by the other: in logical notation,

$$(x)(y)[(x = y) \equiv (F)(Fx \equiv Fy)].$$

This law suggests a way of refuting the identity theory: find some

property that is true of brain states, but not of mental states (or vice versa), and the theory would be exploded.

Spatial properties were often cited to this end. Brain states and processes must of course have some specific spatial location: in the brain as a whole, or in some part of it. And if mental states are identical with brain states, then they must have the very same spatial location. But it is literally meaningless, runs the argument, to say that my feeling-of-pain is located in my ventral thalamus, or that my belief-that-the-sun-is-a-star is located in the temporal lobe of my left cerebral hemisphere. Such claims are as meaningless as the claim that the number 5 is green, or that love weighs twenty grams.

Trying the same move from the other direction, some have argued that it is senseless to ascribe the various *semantic* properties to brain states. Our thoughts and beliefs, for example, have a meaning, a specific propositional content; they are either true or false; and they can enjoy relations such as consistency and entailment. If thoughts and beliefs were brain states, then all these semantic properties would have to be true of brain states. But it is senseless, runs the argument, to say that some resonance in my association cortex is true, or logically entails some other resonance close by, or has the meaning that *P*.

Neither of these moves has the same bite it did twenty years ago, since familiarity with the identity theory and growing awareness of the brain's role have tended to reduce the feelings of semantic oddity produced by the claims at issue. But even if they still struck all of us as semantically confused, this would carry little weight. The claim that sound has a wavelength, or that light has a frequency, must have seemed equally unintelligible in advance of the conviction that both sound and light are wave phenomena. (See, for example, Bishop Berkeley's eighteenth-century dismissal of the idea that sound is a vibratory motion of the air, in Dialogue I of his *Three Dialogues*. The objections are voiced by Philonous.) The claim that warmth is measured in kilogram·meters^2/seconds^2 would have seemed semantically perverse before we understood that temperature is mean molecular kinetic energy. And Copernicus' sixteenth-century claim that the earth *moves* also struck people as absurd to the point of perversity. It is not difficult to appreciate why. Consider the following argument.

> Copernicus' claim that the earth moves is sheer conceptual confusion. For consider what it *means* to say that something moves: "*x* moves" means "*x* changes position relative to the earth." Thus, to say that the earth moves is to say that the earth changes position relative to itself! Which is absurd. Copernicus' position is therefore an abuse of language.

The *meaning analysis* here invoked might well have been correct, but all that would have meant is that the speaker should have set about changing his meanings. The fact is, any language involves a rich network of assumptions about the structure of the world, and if a sentence *S* provokes intuitions of semantic oddness, that is usually because *S* violates one or more of those background assumptions. But one cannot always reject *S* for that reason alone, since the overthrow of those background assumptions may be precisely what the facts require. The 'abuse' of accepted modes of speech is often an essential feature of real scientific progress! Perhaps we shall just have to get used to the idea that mental states have anatomical locations and brain states have semantic properties.

While the charge of sheer senselessness can be put aside, the identity theorist does owe us some account of exactly how physical brain states can have semantic properties. The account currently being explored can be outlined as follows. Let us begin by asking how it is that a particular *sentence* (= utterance type) has the specific propositional content it has: the sentence "La pomme est rouge", for example. Note first that a sentence is always an integrated part of an entire system of sentences: a language. Any given sentence enjoys many relations with countless other sentences: it entails many sentences, is entailed by many others, is consistent with some, is inconsistent with others, provides confirming evidence for yet others, and so forth. And speakers who use that sentence within that language draw inferences in accordance with those relations. Evidently, each sentence (or each set of equivalent sentences) enjoys a unique pattern of such entailment relations: it plays a distinct inferential role in a complex linguistic economy. Accordingly, we say that the sentence "La pomme est rouge" has the propositional content, *the apple is red*, because the sentence "La pomme est rouge" plays *the same role* in French that the sentence "The apple is red" plays in English. To have the relevant propositional content is just to play the relevant inferential role in a cognitive economy.

Returning now to types of brain states, there is no problem in principle in assuming that one's brain is the seat of a complex inferential economy in which types of brain states are the role-playing elements. According to the theory of meaning just sketched, such states would then have propositional content, since having content is not a matter of whether the contentful item is a pattern of sound, a pattern of letters on paper, a set of raised Braille bumps, or a pattern of neural activity. What matters is the inferential role the item plays. Propositional content, therefore, seems within the reach of brain states after all.

We began this subsection with an argument against materialism that appealed to the qualitative *nature* of our mental states, as revealed in

introspection. The next argument appeals to the simple fact that they are introspectible at all.

1. My mental states are introspectively known by me as states of my conscious self.
2. My brain states are *not* introspectively known by me as states of my conscious self.

Therefore, by Leibniz' Law (that numerically identical things must have exactly the same properties),

3. My mental states are not identical with my brain states.

This, in my experience, is the most beguiling form of the argument from introspection, seductive of freshmen and faculty alike. But it is a straightforward instance of a well-known fallacy, which is clearly illustrated in the following parallel arguments:

1. Muhammad Ali is widely known as a heavyweight champion.
2. Cassius Clay is *not* widely known as a heavyweight champion.

Therefore, by Leibniz' Law,

3. Muhammad Ali is not identical with Cassius Clay.

or,

1. Aspirin is recognized by John to be a pain reliever.
2. Acetylsalicylic acid is *not* recogized by John to be a pain reliever.

Therefore, by Leibniz' Law,

3. Aspirin is not identical with acetylsalicylic acid.

Despite the truth of the relevant premises, both conclusions are false: the identities are wholly genuine. Which means that both arguments are invalid. The problem is that the 'property' ascribed in premise (1), and withheld in premise (2), consists only in the subject item's being *recognized, perceived,* or *known* as something-or-other. But such apprehension is not a genuine property of the item itself, fit for divining identities, since one and the same subject may be successfully recognized under one name or description, and yet fail to be recognized under another (accurate, coreferential) description. Bluntly, Leibniz' Law is not valid for these bogus 'properties'. The attempt to use them as above commits what logicians call an *intensional* fallacy. The premises may reflect, not the failure of certain objective identities, but only our continuing failure to appreciate them.

A different version of the preceding argument must also be considered, since it may be urged that one's brain states are more than merely not (yet) known by introspection: they are not know*able* by introspection under any circumstances. Thus,

1. My mental states are knowable by introspection.
2. My brain states are *not* knowable by introspection.

Therefore, by Leibniz' Law,

3. My mental states are not identical with my brain states.

Here the critic will insist that being know*able* by introspection *is* a genuine property of a thing, and that this modified version of the argument is free of the 'intensional fallacy' discussed above.

And so it is. But now the materialist is in a position to insist that the argument contains a false premise—premise (2). For if mental states are indeed brain states, then it is really brain states we have been introspecting all along, though without fully appreciating what they are. And if we can learn to think of and recognize those states under mentalistic descriptions, as we all have, then we can certainly learn to think of and recognize them under their more penetrating neurophysiological descriptions. At the very least, premise (2) simply begs the question against the identity theorist. The mistake is amply illustrated in the following parallel argument:

1. Temperature is knowable by feeling.
2. Mean molecular kinetic energy is *not* knowable by feeling.

Therefore, by Leibniz' Law,

3. Temperature is not identical with mean molecular kinetic energy.

This identity, at least, is long established, and this argument is certainly unsound: premise (2) is false. Just as one can learn to feel that the summer air is about 70°F, or 21°C, so one can learn to feel that the mean KE of its molecules is about 6.2×10^{-21} joules, for whether we realize it or not, that is what our discriminatory mechanisms are keyed to. Perhaps our brain states are similarly accessible. The introspectibility of brain states is addressed again in chapter 8.

Consider now a final argument, again based on the introspectible qualities of our sensations. Imagine a future neuroscientist who comes to know everything there is to know about the physical structure and activity of the brain and its visual system, of its actual and possible states. If for some reason she has never actually *had* a sensation-of-red (because of color blindness, say, or an unusual environment), then there will remain something she does *not* know about certain sensations: *what it is like to have a sensation-of-red.* Therefore, complete knowledge of the physical facts of visual perception and its related brain activity still leaves something out. Accordingly, materialism cannot give an

adequate account of all mental phenomena, and the identity theory must be false.

The identity theorist can reply that this argument exploits an unwitting equivocation on the term "know". Concerning our neuroscientist's utopian knowledge of the brain, "knows" means something like "has mastered the relevant set of neuroscientific propositions". Concerning her (missing) knowledge of what it is like to have a sensation-of-red, "knows" means something like "has a prelinguistic representation of redness in her mechanisms for noninferential discrimination". It is true that one might have the former without the latter, but the materialist is not committed to the idea that having knowledge in the former sense automatically constitutes having knowledge in the second sense. The identity theorist can admit a duality, or even a plurality, of different *types of knowledge* without thereby committing himself to a duality in *types of things known*. The difference between a person who knows all about the visual cortex but has never enjoyed the sensation-of-red, and a person who knows no neuroscience but knows well the sensation-of-red, may reside not in *what* is respectively known by each (brain states by the former, nonphysical *qualia* by the latter), but rather in the different *type*, or *medium*, or *level* of representation each has of exactly the same thing: brain states.

In sum, there are pretty clearly more ways of 'having knowledge' than just having mastered a set of sentences, and the materialist can freely admit that one has 'knowledge' of one's sensations in a way that is independent of the neuroscience one may have learned. Animals, including humans, presumably have a prelinguistic mode of sensory representation. This does not mean that sensations are beyond the reach of physical science. *It just means that the brain uses more modes and media of representation than the mere storage of sentences.* All the identity theorist needs to claim is that those other modes of representation will also yield to neuroscientific explanation.

The identity theory has proved to be very resilient in the face of these predominantly antimaterialist objections. But further objections, rooted in competing forms of materialism, constitute a much more serious threat, as the following sections will show.

Suggested Readings

On the Identity Theory

Feigl, Herbert, "The Mind-Body Problem: *Not* a Pseudo-Problem," in *Dimensions of Mind*, ed. Sidney Hook (New York: New York University Press, 1960).

Place, U. T., "Is Consciousness a Brain Process?" *British Journal of Psychology*, vol. XLVII (1956). Reprinted in *The Philosophy of Mind*, ed. V. C. Chappell (Englewood Cliffs, NJ: Prentice-Hall, 1962).

Smart, J. J. C., "Sensations and Brain Processes," *Philosophical Review*, vol. LXVIII (1959). Reprinted in *The Philosophy of Mind*, ed. V. C. Chappell (Englewood Cliffs, NJ: Prentice-Hall, 1962).

Lewis, David, "An Argument for the Identity Theory," *The Journal of Philosophy*, vol. LXIII, no. 1 (1966).

Nagel, Thomas, "What Is It Like to Be a Bat?" *Philosophical Review*, vol. LXXXIII (1974). Reprinted in *Readings in Philosophy of Psychology*, vol. I, ed. N. Block (Cambridge, MA: Harvard University Press, 1980).

Jackson, Frank, "Epiphenomenal Qualia," *The Philosophical Quarterly*, vol. 32, no. 127 (April, 1982).

On Intertheoretic Reduction

Nagel, Ernst, *The Structure of Science* (New York: Harcourt, Brace, & World, 1961), chapter 11.

Feyerabend, Paul, "Explanation, Reduction, and Empiricism," in *Minnesota Studies in the Philosophy of Science*, vol. III, eds. H. Feigl and G. Maxwell (Minneapolis: University of Minnesota Press, 1962).

Churchland, Paul, *Scientific Realism and the Plasticity of Mind* (Cambridge: Cambridge University Press, 1979), chapter 3, section 11.

Hooker, Clifford, "Towards a General Theory of Reduction," *Dialogue*, vol. XX, nos. 1–3 (1981).

4. *Functionalism*

According to *functionalism*, the essential or defining feature of any type of mental state is the set of causal relations it bears to (1) environmental effects on the body, (2) other types of mental states, and (3) bodily behavior. Pain, for example, characteristically results from some bodily damage or trauma; it causes distress, annoyance, and practical reasoning aimed at relief; and it causes wincing, blanching, and nursing of the traumatized area. Any state that plays exactly that functional role is a pain, according to functionalism. Similarly, other types of mental states (sensations, fears, beliefs, and so on) are also defined by their unique causal roles in a complex economy of internal states mediating sensory inputs and behavioral outputs.

This view may remind the reader of behaviorism, and indeed it is the heir to behaviorism, but there is one fundamental difference between the two theories. Where the behaviorist hoped to define each type of mental state solely in terms of environmental input and behavioral output, the functionalist denies that this is possible. As he sees it, the adequate characterization of almost any mental state involves an ineliminable reference to a variety of other mental states with which it is causally connected, and so a reductive definition solely in terms of publicly observable inputs and outputs is quite impossible. Functionalism is therefore immune to one of the main objections against behaviorism.

Thus the difference between functionalism and behaviorism. The difference between functionalism and the identity theory will emerge from the following argument raised against the identity theory.

Imagine a being from another planet, says the functionalist, a being with an alien physiological constitution, a constitution based on the chemical element silicon, for example, instead of on the element carbon, as ours is. The chemistry and even the physical structure of the alien's brain would have to be systematically different from ours. But even so, that alien brain could well sustain a functional economy of internal states whose mutual *relations* parallel perfectly the mutual relations that define our own mental states. The alien may have an internal state that meets all the conditions for being a pain state, as outlined earlier. That state, considered from a purely physical point of view, would have a very different makeup from a human pain state, but it could nevertheless be identical to a human pain state from a purely functional point of view. And so for all of his functional states.

If the alien's functional economy of internal states were indeed *functionally isomorphic* with our own internal economy—if those states were causally connected to inputs, to one another, and to behavior in ways that parallel our own internal connections—then the alien would have pains, and desires, and hopes, and fears just as fully as we, despite the differences in the physical system that sustains or realizes those functional states. What is important for mentality is not the matter of which the creature is made, but the structure of the internal activities which that matter sustains.

If we can think of one alien constitution, we can think of many, and the point just made can also be made with an artificial system. Were we to create an electronic system—a computer of some kind—whose internal economy were functionally isomorphic with our own in all the relevant ways, then it too would be the subject of mental states.

What this illustrates is that there are almost certainly many more ways than one for nature, and perhaps even for man, to put together a thinking, feeling, perceiving creature. And this raises a problem for the identity theory, for it seems that there is no single type of physical state to which a given type of mental state must always correspond. Ironically, there are *too many* different kinds of physical systems that can realize the functional economy characteristic of conscious intelligence. If we consider the universe at large, therefore, and the future as well as the present, it seems quite unlikely that the identity theorist is going to find the one-to-one match-ups between the concepts of our common-sense mental taxonomy and the concepts of an overarching theory that encompasses all of the relevant physical systems. But that is what intertheoretic reduction is standardly said to require. The prospects for universal identities, between types of mental states and types of brain states, are therefore slim.

If the functionalists reject the traditional 'mental-type = physical type' identity theory, virtually all of them remain committed to a weaker 'mental token = physical token' identity theory, for they still maintain that each *instance* of a given type of mental state is numerically identical with some specific physical state in some physical system or other. It is only universal (type/type) identities that are rejected. Even so, this rejection is typically taken to support the claim that the science of psychology is or should be *methodologically autonomous* from the various physical sciences such as physics, biology, and even neurophysiology. Psychology, it is claimed, has its own irreducible laws and its own abstract subject matter.

As this book is written, functionalism is probably the most widely held theory of mind among philosophers, cognitive psychologists, and artificial intelligence researchers. Some of the reasons are apparent from the preceding discussion, and there are further reasons as well. In characterizing mental states as essentially functional states, functionalism places the concerns of psychology at a level that abstracts

from the teeming detail of a brain's neurophysiological (or crystallographic, or microelectronic) structure. The science of psychology, it is occasionally said, is methodologically autonomous from those other sciences (biology, neuroscience, circuit theory) whose concerns are with what amount to engineering details. This provides a rationale for a great deal of work in cognitive psychology and artificial intelligence, where researchers postulate a system of abstract functional states and then test the postulated system, often by way of its computer simulation, against human behavior in similar circumstances. The aim of such work is to discover in detail the functional organization that makes us what we are. (Partly in order to evaluate the prospects for a functionalist philosophy of mind, we shall examine some of the recent research in artificial intelligence in chapter 6.)

Arguments against Functionalism

Current popularity aside, functionalism also faces difficulties. The most commonly posed objection cites an old friend: sensory qualia. Functionalism may escape one of behaviorism's fatal flaws, it is said, but it still falls prey to the other. By attempting to make its *relational* properties the definitive feature of any mental state, functionalism ignores the 'inner' or qualitative nature of our mental states. But their qualitative nature is the essential feature of a great many types of mental state (pain, sensations of color, of temperature, of pitch, and so on), runs the objection, and functionalism is therefore false.

The standard illustration of this apparent failing is called "the inverted spectrum thought-experiment". It is entirely conceivable, runs the story, that the range of color sensations that I enjoy upon viewing standard objects is simply inverted relative to the color sensations that you enjoy. When viewing a tomato, I may have what is really a sensation-of-green where you have the normal sensation-of-red; when viewing a banana, I may have what is really sensation-of-blue where you have the normal sensation-of-yellow; and so forth. But since we have no way of comparing our inner qualia, and since I shall make all the same observational discriminations among objects that you will, there is no way to tell whether my spectrum is inverted relative to yours.

The problem for functionalism arises as follows. Even if my spectrum is inverted relative to yours, we remain functionally isomorphic with one another. My visual sensation upon viewing a tomato is *functionally* identical with your visual sensation upon viewing a tomato. According to functionalism, therefore, they are the very same type of state, and it does not even make sense to suppose that my sensation is 'really' a sensation-of-green. If it meets the functional conditions for being a sensation-of-red, then by definition it is a sensation-of-red. According to functionalism, apparently, a spectrum inversion of the kind described is ruled out by definition. But such inversions are entirely conceivable, concludes the objection, and if functionalism entails that they are not conceivable, then functionalism is false.

Another qualia-related worry for functionalism is the so-called "absent qualia problem". The functional organization characteristic of conscious intelligence can be instantiated (= realized or instanced) in a considerable variety of physical systems, some of them radically different from a normal human. For example, a giant electronic computer might instantiate it, and there are more radical possibilities still. One writer asks us to imagine the people of China—all 10^9 of them—organized into an intricate game of mutual interactions so that collectively they constitute a giant brain which exchanges inputs and outputs with a single robot body. That system of the robot-plus-10^9-unit-brain could presumably instantiate the relevant functional organization (though no doubt it would be much slower in its activities than a human or a computer), and would therefore be the subject of mental states, according to functionalism. But surely, it is urged, the complex states that there play the functional roles of pain, pleasure, and sensations-of-color would not have intrinsic qualia as ours do, and would therefore fail to be genuine mental states. Again, functionalism seems at best an incomplete account of the nature of mental states.

It has recently been argued that both the inverted-qualia and the absent-qualia objections can be met, without violence to functionalism and without significant violence to our common-sense intuitions about qualia. Consider the inversion problem first. I think the functionalist is right to insist that the type-identity of our visual sensations be reckoned according to their functional role. But the objector is also right in insisting that a relative inversion of two people's qualia, without functional inversion, is entirely conceivable. The apparent inconsistency between these positions can be dissolved by insisting that (1) our functional states (or rather, their physical realizations) do indeed have an intrinsic nature on which our introspective identification of those states depends; while also insisting that (2) such intrinsic natures are nevertheless not essential to the type-identity of a given mental state, and may indeed *vary* from instance to instance of the same type of mental state.

What this means is that the qualitative character of your sensation-of-red might be different from the qualitative character of my sensation-of-red, slightly or substantially, and a third person's sensation-of-red might be different again. But so long as all three states are standardly caused by red objects and standardly cause all three of us to believe

that something is red, then all three states are sensations-of-red, whatever their intrinsic qualitative character. Such intrinsic qualia merely serve as salient features that permit the quick introspective identification of sensations, as black-on-orange stripes serve as a salient feature for the quick visual identification of tigers. But specific qualia are not essential to the type-identity of mental states, any more than black-on-orange stripes are essential to the type-identity of tigers.

Plainly, this solution requires the functionalist to admit the *reality* of qualia, and we may wonder how there can be room for qualia in his materialist world-picture. Perhaps they can be fit in as follows: *identify* them with physical properties of whatever physical states instantiate the mental (functional) states that display them. For example, identify the qualitative nature of your sensations-of-red with that physical feature (of the brain state that instantiates it) to which your mechanisms of introspective discrimination are in fact responding when you judge that you have a sensation-of-red. If materialism is true, then there must *be* some internal physical feature or other to which your discrimination of sensations-of-red is keyed: *that* is the quale of your sensations-of-red. If the pitch of a sound can turn out to be the frequency of an oscillation in air pressure, there is no reason why the quale of a sensation cannot turn out to be, say, a spiking frequency in a certain neural pathway. ('Spikes' are the tiny electrochemical pulses by which our brain cells communicate.)

This entails that creatures with a constitution different from ours may have qualia different from ours, despite being psychologically isomorphic with us. It does not entail that they *must* have different qualia, however. If the qualitative character of my sensation-of-red is really a spiking frequency of 90 hertz in a certain neural pathway, it is possible that an electromechanical robot might enjoy the very same qualitative character if, in reporting sensations-of-red, the robot were responding to a spiking frequency of 90 hertz in a corresponding *copper* pathway. It might be the spiking frequency that matters to our respective mechanisms of discrimination, not the nature of the medium that carries it.

This proposal also suggests a solution to the absent qualia problem. So long as the physical system at issue is functionally isomorphic with us, to the last detail, then it will be equally capable of subtle introspective discriminations among its sensations. Those discriminations must be made on some systematic physical basis, that is, on some characteristic physical features of the states being discriminated. Those features at the objective focus of the system's discriminatory mechanisms, *those* are its sensory qualia—though the alien system is no more likely to appreciate their true physical nature than we appeciate the true physical

nature of our own qualia. Sensory qualia are therefore an inevitable concomitant of any system with the kind of functional organization at issue. It may be difficult or impossible to 'see' the qualia in an alien system, but it is equally difficult to 'see' them even when looking into a human brain.

I leave it to the reader to judge the adequacy of these responses. If they are adequate, then, given its other virtues, functionalism must be conceded a very strong position among the competing contemporary theories of mind. It is interesting, however, that the defense offered in the last paragraph found it necessary to take a leaf from the identity theorist's book (types of quale are reduced to or identified with types of physical state), since the final objection we shall consider also tends to blur the distinction between functionalism and reductive materialism.

Consider the property of *temperature*, runs the objection. Here we have a paradigm of a physical property, one that has also been cited as the paradigm of a successfully *reduced* property, as expressed in the intertheoretic identity

> "temperature = mean kinetic energy of constituent molecules".

Strictly speaking, however, this identity is true only for the temperature of a gas, where simple particles are free to move in ballistic fashion. In a *solid*, temperature is realized differently, since the interconnected molecules are confined to a variety of vibrational motions. In a *plasma*, temperature is something else again, since a plasma has no constituent molecules; they, and their constituent atoms, have been ripped to pieces. And even a *vacuum* has a so-called 'blackbody' temperature—in the distribution of electromagnetic waves coursing through it. Here temperature has nothing to do with the kinetic energy of particles.

It is plain that the physical property of temperature enjoys 'multiple instantiations' no less than do psychological properties. Does this mean that thermodynamics (the theory of heat and temperature) is an 'autonomous science', separable from the rest of physics, with its own irreducible laws and its own abstract nonphysical subject matter?

Presumably not. What it means, concludes the objection, is that *reductions are domain-specific*:

> temperature-in-a-gas = the mean kinetic energy of the gas's molecules,

whereas

> temperature-in-a-vacuum = the blackbody distribution of the vacuum's transient radiation.

Similarly, perhaps

joy-in-a-human = resonances in the lateral hypothalamus,

whereas

joy-in-a-Martian = something else entirely.

This means that we may expect some type/type reductions of mental states to physical states after all, though they will be much narrower than was first suggested. Furthermore, it means that functionalist claims concerning the radical autonomy of psychology cannot be sustained. And last, it suggests that functionalism is not so profoundly different from the identity theory as was first made out.

As with the defense of functionalism outlined earlier, I leave the evaluation of this criticism to the reader. We shall have occasion for further discussion of functionalism in later chapters. At this point, let us turn to the final materialist theory of mind, for functionalism is not the only major reaction against the identity theory.

Suggested Readings

Putnam, Hilary, "Minds and Machines," in *Dimensions of Mind*, ed. Sidney Hook (New York: New York University Press, 1960).

Putnam, Hilary, "Robots: Machines or Artificially Created Life?" *Journal of Philosophy*, vol. LXI, no. 21 (1964).

Putnam, Hilary, "The Nature of Mental States," in *Materialism and the Mind-Body Problem*, ed. David Rosenthal (Englewood Cliffs, NJ: Prentice-Hall, 1971).

Fodor, Jerry, *Psychological Explanation* (New York: Random House, 1968).

Dennett, Daniel, *Brainstorms* (Montgomery, Vermont: Bradford, 1978).

Concerning Difficulties with Functionalism

Block, Ned, "Troubles with Functionalism," in *Minnesota Studies in the Philosophy of Science*, vol. IX ed. C. W. Savage (Minneapolis: University of Minnesota Press, 1978). Reprinted in *Readings in Philosophy of Psychology*, ed. N. Block (Cambridge, MA: Harvard University Press, 1980).

Churchland, Paul and Patricia, "Functionalism, Qualia, and Intentionality," *Philosophical Topics*, vol. 12, no. 1 (1981). Reprinted in *Mind, Brain, and Function*, eds. J. Biro and R. Shahan (Norman, OK: University of Oklahoma Press, 1982).

Churchland, Paul, "Eliminative Materialism and the Propositional Attitudes," *Journal of Philosophy*, vol. LXXVIII, no. 2 (1981).

Shoemaker, Sidney, "The Inverted Spectrum," *Journal of Philosophy*, vol. LXXIX, no. 7 (1982).

Enc, Berent, "In Defense of the Identity Theory," *Journal of Philosophy*, vol. LXXX, no. 5 (1983).

5. Eliminative Materialism

The identity theory was called into doubt not because the prospects for a materialist account of our mental capacities were thought to be poor, but because it seemed unlikely that the arrival of an adequate materialist theory would bring with it the nice one-to-one match-ups, between the concepts of folk psychology and the concepts of theoretical neuroscience, that intertheoretic reduction requires. The reason for that doubt was the great variety of quite different physical systems that could instantiate the required functional organization. *Eliminative materialism* also doubts that the correct neuroscientific account of human capacities will produce a neat reduction of our common-sense framework, but here the doubts arise from a quite different source.

As the eliminative materialists see it, the one-to-one match-ups will not be found, and our common-sense psychological framework will not enjoy an intertheoretic reduction, *because our common-sense psychological framework is a false and radically misleading conception of the causes of human behavior and the nature of cognitive activity*. On this view, folk psychology is not just an incomplete representation of our inner natures; it is an outright *mis*representation of our internal states and activities. Consequently, we cannot expect a truly adequate neuroscientific account of our inner lives to provide theoretical categories that match up nicely with the categories of our common-sense framework. Accordingly, we must expect that the older framework will simply be eliminated, rather than be reduced, by a matured neuroscience.

Historical Parallels As the identity theorist can point to historical cases of successful intertheoretic reduction, so the eliminative materialist can point to historical cases of the outright elimination of the ontology of an older theory in favor of the ontology of a new and superior theory. For most of the eighteenth and nineteenth centuries, learned people believed that heat was a subtle *fluid* held in bodies, much in the way water is held in a sponge. A fair body of moderately successful theory described the way this fluid substance—called "caloric"—flowed within a body, or from one body to another, and how it produced thermal expansion, melting, boiling, and so forth. But by the end of the last century it had become abundantly clear that heat was not a substance at all, but just the energy of motion of the trillions of jostling molecules that make up the heated body itself. The new theory—the "corpuscular/kinetic theory of matter and heat"—was much more successful than the old in explaining and predicting the thermal behavior of bodies. And since we were unable to *identify* caloric fluid with kinetic energy (according to the old theory,

caloric is a material *substance;* according to the new theory, kinetic energy is a form of *motion*), it was finally agreed that there is *no such thing* as caloric. Caloric was simply eliminated from our accepted ontology.

A second example. It used to be thought that when a piece of wood burns, or a piece of metal rusts, a spiritlike substance called "phlogiston" was being released: briskly, in the former case, slowly in the latter. Once gone, that 'noble' substance left only a base pile of ash or rust. It later came to be appreciated that both processes involve, not the loss of something, but the *gaining* of a substance taken from the atmosphere: oxygen. Phlogiston emerged, not as an incomplete description of what was going on, but as a radical misdescription. Phlogiston was therefore not suitable for reduction to or identification with some notion from within the new oxygen chemistry, and it was simply eliminated from science.

Admittedly, both of these examples concern the elimination of something nonobservable, but our history also includes the elimination of certain widely accepted 'observables'. Before Copernicus' views became available, almost any human who ventured out at night could look up at *the starry sphere of the heavens,* and if he stayed for more than a few minutes he could also see that it *turned,* around an axis through Polaris. What the sphere was made of (crystal?) and what made it turn (the gods?) were theoretical questions that exercised us for over two millennia. But hardly anyone doubted the existence of what everyone could observe with their own eyes. In the end, however, we learned to reinterpret our visual experience of the night sky within a very different conceptual framework, and the turning sphere evaporated.

Witches provide another example. Psychosis is a fairly common affliction among humans, and in earlier centuries its victims were standardly seen as cases of demonic possession, as instances of Satan's spirit itself, glaring malevolently out at us from behind the victims' eyes. That witches exist was not a matter of any controversy. One would occasionally see them, in any city or hamlet, engaged in incoherent, paranoid, or even murderous behavior. But observable or not, we eventually decided that witches simply do not exist. We concluded that the concept of a witch is an element in a conceptual framework that misrepresents so badly the phenomena to which it was standardly applied that literal application of the notion should be permanently withdrawn. Modern theories of mental dysfunction led to the elimination of witches from our serious ontology.

The concepts of folk psychology—belief, desire, fear, sensation, pain, joy, and so on—await a similar fate, according to the view at issue. And when neuroscience has matured to the point where the poverty

of our current conceptions is apparent to everyone, and the superiority of the new framework is established, we shall then be able to set about *re*conceiving our internal states and activities, within a truly adequate conceptual framework at last. Our explanations of one another's behavior will appeal to such things as our neuropharmacological states, the neural activity in specialized anatomical areas, and whatever other states are deemed relevant by the new theory. Our private introspection will also be transformed, and may be profoundly enhanced by reason of the more accurate and penetrating framework it will have to work with—just as the astronomer's perception of the night sky is much enhanced by the detailed knowledge of modern astronomical theory that he or she possesses.

The magnitude of the conceptual revolution here suggested should not be minimized: it would be enormous. And the benefits to humanity might be equally great. If each of us possessed an accurate neuroscientific understanding of (what we now conceive dimly as) the varieties and causes of mental illness, the factors involved in learning, the neural basis of emotions, intelligence, and socialization, then the sum total of human misery might be much reduced. The simple increase in mutual understanding that the new framework made possible could contribute substantially toward a more peaceful and humane society. Of course, there would be dangers as well: increased knowledge means increased power, and power can always be misused.

Arguments for Eliminative Materialism

The arguments for eliminative materialism are diffuse and less than decisive, but they are stronger than is widely supposed. The distinguishing feature of this position is its denial that a smooth intertheoretic reduction is to be expected—even a species-specific reduction—of the framework of folk psychology to the framework of a matured neuroscience. The reason for this denial is the eliminative materialist's conviction that folk psychology is a hopelessly primitive and deeply confused conception of our internal activities. But why this low opinion of our common-sense conceptions?

There are at least three reasons. First, the eliminative materialist will point to the widespread explanatory, predictive, and manipulative failures of folk psychology. So much of what is central and familiar to us remains a complete mystery from within folk psychology. We do not know what *sleep* is, or why we have to have it, despite spending a full third of our lives in that condition. (The answer, "For rest," is mistaken. Even if people are allowed to rest continuously, their need for sleep is undiminished. Apparently, sleep serves some deeper functions, but we do not yet know what they are.) We do not understand how *learning*

transforms each of us from a gaping infant to a cunning adult, or how differences in *intelligence* are grounded. We have not the slightest idea how *memory* works, or how we manage to retrieve relevant bits of information instantly from the awesome mass we have stored. We do not know what *mental illness* is, nor how to cure it.

In sum, the most central things about us remain almost entirely mysterious from within folk psychology. And the defects noted cannot be blamed on inadequate time allowed for their correction, for folk psychology has enjoyed no significant changes or advances in well over 2,000 years, despite its manifest failures. Truly successful theories may be expected to reduce, but significantly unsuccessful theories merit no such expectation.

This argument from explanatory poverty has a further aspect. So long as one sticks to normal brains, the poverty of folk psychology is perhaps not strikingly evident. But as soon as one examines the many perplexing behavioral and cognitive deficits suffered by people with *damaged* brains, one's descriptive and explanatory resources start to claw the air (see, for example chapter 7.3, p. 143). As with other humble theories asked to operate successfully in unexplored extensions of their old domain (for example, Newtonian mechanics in the domain of velocities close to the velocity of light, and the classical gas law in the domain of high pressures or temperatures), the descriptive and explanatory inadequacies of folk psychology become starkly evident.

The second argument tries to draw an inductive lesson from our conceptual history. Our early folk theories of motion were profoundly confused, and were eventually displaced entirely by more sophisticated theories. Our early folk theories of the structure and activity of the heavens were wildly off the mark, and survive only as historical lessons in how wrong we can be. Our folk theories of the nature of fire, and the nature of life, were similarly cockeyed. And one could go on, since the vast majority of our past folk conceptions have been similarly exploded. All except folk psychology, which survives to this day and has only recently begun to feel pressure. But the phenomenon of conscious intelligence is surely a more complex and difficult phenomenon than any of those just listed. So far as accurate understanding is concerned, it would be a *miracle* if we had got *that* one right the very first time, when we fell down so badly on all the others. Folk psychology has survived for so very long, presumably, not because it is basically correct in its representations, but because the phenomena addressed are so surpassingly difficult that any useful handle on them, no matter how feeble, is unlikely to be displaced in a hurry.

A third argument attempts to find an a priori advantage for eliminative materialism over the identity theory and functionalism. It attempts to

counter the common intuition that eliminative materialism is distantly possible, perhaps, but is much less probable than either the identity theory or functionalism. The focus again is on whether the concepts of folk psychology will find vindicating match-ups in a matured neuroscience. The eliminativist bets no; the other two bet yes. (Even the functionalist bets yes, but expects the match-ups to be only species-specific, or only person-specific. Functionalism, recall, denies the existence only of *universal* type/type identities.)

The eliminativist will point out that the requirements on a reduction are rather demanding. The new theory must entail a set of principles and embedded concepts that mirrors very closely the specific conceptual structure to be reduced. And the fact is, there are vastly many more ways of being an explanatorily successful neuroscience while *not* mirroring the structure of folk psychology, than there are ways of being an explanatorily successful neuroscience while also *mirroring* the very specific structure of folk psychology. Accordingly, the a priori probability of eliminative materialism is not lower, but substantially *higher* than that of either of its competitors. One's initial intuitions here are simply mistaken.

Granted, this initial a priori advantage could be reduced if there were a very strong presumption in favor of the truth of folk psychology—true theories are better bets to win reduction. But according to the first two arguments, the presumptions on this point should run in precisely the opposite direction.

Arguments against Eliminative Materialism

The initial plausibility of this rather radical view is low for almost everyone, since it denies deeply entrenched assumptions. That is at best a question-begging complaint, of course, since those assumptions are precisely what is at issue. But the following line of thought does attempt to mount a real argument.

Eliminative materialism is false, runs the argument, because one's introspection reveals directly the existence of pains, beliefs, desires, fears, and so forth. Their existence is as obvious as anything could be.

The eliminative materialist will reply that this argument makes the same mistake that an ancient or medieval person would be making if he insisted that he could just see with his own eyes that the heavens form a turning sphere, or that witches exist. The fact is, all observation occurs within some system of concepts, and our observation judgments are only as good as the conceptual framework in which they are expressed. In all three cases—the starry sphere, witches, and the familiar mental states—precisely what is challenged is the integrity of the background conceptual frameworks in which the observation judgments

are expressed. To insist on the validity of one's experiences, *traditionally interpreted,* is therefore to beg the very question at issue. For in all three cases, the question is whether we should *re*conceive the nature of some familiar observational domain.

A second criticism attempts to find an incoherence in the eliminative materialist's position. The bald statement of eliminative materialism is that the familiar mental states do not really exist. But that statement is meaningful, runs the argument, only if it is the expression of a certain *belief,* and an *intention* to communicate, and a *knowledge* of the language, and so forth. But if the statement is true, then no such mental states exist, and the statement is therefore a meaningless string of marks or noises, and cannot be true. Evidently, the assumption that eliminative materialism is true entails that it cannot be true.

The hole in this argument is the premise concerning the conditions necessary for a statement to be meaningful. It begs the question. If eliminative materialism is true, then meaningfulness must have some different source. To insist on the 'old' source is to insist on the validity of the very framework at issue. Again, an historical parallel may be helpful here. Consider the medieval theory that being biologically *alive* is a matter of being ensouled by an immaterial *vital spirit.* And consider the following response to someone who has expressed disbelief in that theory.

> My learned friend has stated that there is no such thing as vital spirit. But this statement is incoherent. For if it is true, then my friend does not have vital spirit, and must therefore be *dead.* But if he is dead, then his statement is just a string of noises, devoid of meaning or truth. Evidently, the assumption that antivitalism is true entails that it cannot be true! Q.E.D.

This second argument is now a joke, but the first argument begs the question in exactly the same way.

A final criticism draws a much weaker conclusion, but makes a rather stronger case. Eliminative materialism, it has been said, is making mountains out of molehills. It exaggerates the defects in folk psychology, and underplays its real successes. Perhaps the arrival of a matured neuroscience will require the elimination of the occasional folk-psychological concept, continues the criticism, and a minor adjustment in certain folk-psychological principles may have to be endured. But the large-scale elimination forecast by the eliminative materialist is just an alarmist worry or a romantic enthusiasm.

Perhaps this complaint is correct. And perhaps it is merely complacent. Whichever, it does bring out the important point that we do not confront two simple and mutually exclusive possibilities here: pure reduction

versus pure elimination. Rather, these are the end points of a smooth spectrum of possible outcomes, between which there are mixed cases of partial elimination and partial reduction. Only empirical research (see chapter 7) can tell us where on that spectrum our own case will fall. Perhaps we should speak here, more liberally, of "revisionary materialism", instead of concentrating on the more radical possibility of an across-the-board elimination. Perhaps we should. But it has been my aim in this section to make it at least intelligible to you that our collective conceptual destiny lies substantially toward the revolutionary end of the spectrum.

Suggested Readings

Feyerabend, Paul, "Comment: 'Mental Events and the Brain,' " *Journal of Philosophy,* vol. LX (1963). Reprinted in *The Mind/Brain Identity Theory,* ed. C. V. Borst (London: Macmillan, 1970).

Feyerabend, Paul, "Materialism and the Mind-Body Problem," *Review of Metaphysics,* vol. XVII (1963). Reprinted in *The Mind/Brain Identity Theory,* ed. C. V. Borst (London: Macmillan, 1970).

Rorty, Richard, "Mind-Body Identity, Privacy, and Categories," *Review of Metaphysics,* vol. XIX (1965). Reprinted in *Materialism and the Mind-Body Problem,* ed. D. M. Rosenthal (Englewood Cliffs, NJ: Prentice-Hall, 1971).

Rorty, Richard, "In Defense of Eliminative Materialism," *Review of Metaphysics,* vol. XXIV (1970). Reprinted in *Materialism and the Mind-Body Problem,* ed. D. M. Rosenthal (Englewood Cliffs, NJ: Prentice-Hall, 1971).

Churchland, Paul, "Eliminative Materialism and the Propositional Attitudes," *Journal of Philosophy,* vol. LXXVIII, no. 2 (1981).

Dennett, Daniel, "Why You Can't Make a Computer that Feels Pain," in *Brainstorms* (Montgomery, VT: Bradford, 1978).