1 Minds, Brains, and Behavior

It is usually assumed that when we say "the mind is the brain" we are taking a concept from neurophysiology (brain), and saying that it translates to a concept from common sense (mind). In fact, something very much like the opposite is the case. The idea that the brain is the organ of the body that feels and thinks was not something discovered by modern science. It is at least 2,000 years old, for Hippocrates wrote "Men ought to know that from the brain, and the brain alone, arise our pleasures, joys, laughter, and jests, as well as our sorrows, joys, and fears" (quoted in Bailey 1975, p. 10). Plato also believed that reason resided in the brain, although he thought that courage and ambition resided in the heart, and desire resided in the stomach. When Aristotle said the heart was the center of the soul, and that the brain's function was to cool the blood, he was probably contradicting the common wisdom of the time, not stating it. It is thus not surprising that ordinary language is filled with assumptions that the mind is the brain. After all, people do speak of thinking as "using your brain," of stupid people as being "brainless," and so on. Neuroscientists (like everyone else) learned the mind–brain identity at their mothers' knee and brought it with them to the data.

Yet although there is this vague sense that the brain is somehow responsible for mental activities and phenomena, there is no clear understanding in common sense of *how* the brain is so responsible. So when Gilbert Ryle began to explicate a commonsense alternative to dualism in his *The Concept of Mind*, he made almost no reference to the brain at all. Instead he tried to explain the ordinary concept of mind in terms of what ordinary people experience: human behavior, both their own and other people's. He claimed that when we say "Jones is in pain," what we mean is that Jones is either wincing, or jumping up and down and holding his

thumb, or doing some other combination of behaviors that we have learned to associate with pain. Ryle claimed that when I say "I am in pain," what I mean is either that I observe myself performing such behaviors, or that I experience a disposition to perform such behaviors, which I must repress. Similarly, statements like "I believe that it will rain this afternoon" supposedly could be replaced in principle by lists of statements about behavior (and disposition to behave) that made no reference to beliefs or other mental entities. Unfortunately, it was simply impossible to describe mental states by substituting descriptions that referred only to behavior. The alleged substitutions turned out to be infinitely long, and/or to have other statements in them that referred to beliefs. Mental entities were simply too tough to yield to Occam's razor, so philosophers had to accept the fact that in some sense there really are such things as thoughts, beliefs, pains, and pleasures. But because no one wanted to return to Cartesian dualism, these had to be some sort of physical things. But what?

The next answer, suggested by D. M. Armstrong, J. J. C. Smart, and U. T. Place was that mental states are really brain states. Being philosophers, however, they were not claiming to have proven this in the laboratory. Their point was somewhat similar to one made by Hilary Putnam several years later (Putnam 1975). Many concepts in ordinary language are considered to be understood if you know which experts to ask for further clarification. If the ancient Greeks ever talked about brain states, they might have meant "those occurrences in the skull that Hippocrates could explain to us if we asked him" and the modern person-on-the-street's understanding of "brain state" is essentially the same, except that we have better-informed experts to fall back on. Thus when the mind–brain identity theorists tried to explicate the meaning of commonsense concepts about mentality, they used carefully ambiguous locutions like "something is going on in me which is like what goes on when I have my eyes open and there is an orange in front of me." Or else they compared references to brain states to phrases like "someone telephoned" in which we later identify who the someone is (Borst 1970, pp. 14, 28). The mind–brain identity theorists believed that it is the job of neuroscientists to identify the "something" we are referring to when we say I am having a sensation of red, or a thought about George Washington. Because discourse about brains appears to be very different from discourse about feelings and

thoughts, the mind–body problem came to be seen as a subspecies of the sense–reference problem. If *X* is both a brain state and a thought, isn't this the same situation as when *X* is both the morning star and the evening star, or both the mayor of Dublin and the ugliest Irishman?

There are numerous problems with this position, however, and two closely related alternatives arose to circumvent them and to create new problems of their own.

Functionalism

Functionalism pointed out that there are too many different physical ways that mental predicates could be instantiated for them to be reduced to single physical predicates. Nor was this problem limited to the mind–brain relationship. It arose in biology, economics, and almost any science other than physics when one tried to identify its kinds and predicates with purely physical terms. As Fodor points out, a monetary exchange could be instantiated physically by handing over a dollar bill, or by writing a check, or by using a string of wampum, and it would obviously be only an improbable coincidence if any of these actions had anything in common physically (Fodor 1975, ch. 1). And as Putnam (1960) pointed out, the same problem arises when we talk about the physical substrate that instantiates beliefs or pains or fears in humans, dogs, and Martians. Even if it were a matter of empirical fact that there was some physical attribute they all had in common (perhaps they are all made of protein), this would be a trivial coincidence that would probably tell us nothing of importance, and certainly leave out much that was essential. The factor that makes a belief a belief, or a sensation a sensation, is not what it is made of, but the functional role it performs in a biological and cognitive system.

Like any philosophical position that survives for any length of time, functionalism has received a lot of criticism. It has been taken by Ned Block (Block 1978) to be incapable of accounting for subjective experience (although Block has admitted in conversation that he still considers himself to be a functionalist in some sense). Fodor has implied that functionalism provides proof that psychology can be an autonomous science, a position that the Churchlands have, in my view, successfully refuted (especially Churchland 1989, pp. 12–17). Putnam (1988) has renounced his own version of functionalism, because it ignored multiple realizability

at the computational level while acknowledging that it existed between the computational and the physical. But none of these criticisms has shaken the fundamental insight of functionalism: that physical kinds cannot be the only kinds in the world, and therefore the language of physics cannot tell the whole story about the way things are. Physics will always have *something* to say about everything we encounter. Even though functionalism denies the existence of type–type identities between the physical and the functional, each token of any functional type is physical, and therefore functionalism is usually considered to be a kind of physicalism. But the fact of multiple realizability guarantees that the physical story cannot be the whole story.

Functionalism poses a more serious threat to the mind–brain identity theory than is usually acknowledged. In many ways, it is a revitalized form of Rylean behaviorism, for it defines the mind in terms of what we do rather than what we are made of. But unlike behaviorism, functionalism grants a genuine ontological independence to mental entities, an independence that apparently frees them not only from behavior, but also from brains. It therefore leaves open at least the possibility that whatever replaces the concept of mind might not be a precisely bordered chunk of biological stuff. If the mind is seen as identical with certain abstract causal roles performed by an organism or its parts, almost any part of the body could be seen as mental when it performed those roles, and some such roles might even be performed by the entire body (the way moves in chess are performed by an entire chess piece). If this were the case, no part of the body would be identical with the mind, just as no single building would be identical with Oxford University.

According to functionalism, the physical characteristics of the brain embody the mind, but they are not essential to the nature of mind. Consequently, many people used functionalism as a way of freeing the study of the mind from the study of the brain. However, at the same time that functionalism was formulated, another philosophical position called eliminative materialism was demanding that brain studies be the sole, or at least the primary, source of information about the mind.

Eliminative Materialism

Eliminative materialism was first formulated by Richard Rorty and Paul Feyerabend (see their articles in Borst 1970 and Rosenthal 1971) and then

developed into a manifesto for a research program by Paul Churchland. The eliminative materialists claimed that the problems of the mind–brain identity thesis can be dissolved by simply saying that there will be no one-to-one correspondence in future neuroscience between mental events and physical events. In fact, they claim that future neuroscience may prove that there are no such entities as thoughts or sensations, and never were. The fact that functional states cannot be identified with brain states does not necessarily show that they have an independent reality distinct from brain states. The history of science has shown us that when a scientific reduction takes place, it is often impossible to formulate what were called bridge laws, that is, logical identities between entities in the reduced and reducing domains. But this does not make the entities in the old theory independent, it makes them nonexistent. We did not establish identities between the chemical elements and the alchemical essences. Why should we assume that we can establish identities between mental states and brain states?

The fullest articulation of this position is in *Eliminative Materialism and Propositional Attitudes* (P. M. Churchland 1989, pp. 1–22), where Paul Churchland claims that "mind" and all of those entities that allegedly inhabit mind, such as beliefs, hopes, sensations, thoughts, and so on, are part of a conceptual system he calls *folk psychology*. Churchland also claims that folk psychology does not have any claims to certainty, that Descartes was wrong when he said that direct introspection could produce an infallible awareness of the mind and its contents. Because folk psychology is based on personal introspection, not laboratory research, it could be just plain wrong about many things, just as folk physics was wrong when it claimed that heavy objects fall faster than light ones, and that the earth is flat. We should therefore be willing to look at research on brains as *the* source of new information about our minds, and whenever this research contradicts our commonsense view of ourselves, we should be willing to accept that the brain researchers are right and that common sense is wrong.

Eliminative materialism and functionalism have no official quarrel with each other, although each has been unjustifiably pressed into the service of other causes that have created the illusion of conflict. The essential points that both sides agree on can, I think, be summed up in the following four principles. In fact I don't see how anyone who has faith in the scientific method could doubt these principles. If I were not a pragmatist, I would probably call them "fundamental a priori principles" or

"necessary truths." But although I recognize that these presuppositions are doubtable in principle, I accept Peirce's maxim that we must not doubt in our philosophy what we do not doubt in our hearts. Few people in the cognitive science community would question these four claims, and the rest of this book will be written on the assumption that they are true.

(1) Mental properties are not inherent in some particular physical stuff. This was not always as obvious as it is now. In William James's time, there were scientists who claimed that thought was phosphorus, because they found large amounts of this element in the brain. This was what gave rise to the modern folk idea that fish is brain food. And from James's description (James 1890, vol. 1, p. 101), these scientists appeared to believe that this relationship was a straightforward identity, as if a bottle of phosphorus sitting on a chemist's shelf would be vaguely thinking about something or other. John Searle occasionally appears to be advocating a similar position when he claims that consciousness is a biological, not a functional property, and infers from this that consciousness cannot be any sort of abstract pattern. But as arguments by Lycan (1987) and Millikan (1984) show, biological properties are also functional properties. A heart is a heart because it performs a particular function in the circulatory system, not because it is a particular shape or made out of a particular kind of protein. The shape and chemical composition of any particular heart will determine its ability to perform its function. But that is because those physical characteristics must relate to other physical characteristics of other parts of the system. It is not an intrinsic or necessary characteristic of all hearts that they must be a particular shape or substance.

(2) This therefore means that mentality must be a property of some kind of system. This system must consist of parts,[1] each of which must have certain physical characteristics within the context of that particular system. These physical characteristics are constitutive of consciousness only within the context of that system, however. In and of themselves, no particular physical characteristic is essential to consciousness. This is the fundamental assumption of what is called strong AI: that it is possible in principle to build a conscious creature out of silicon, even though all such creatures we know of are made of protein. Silicon could turn out, for some physical reason, not to be flexible enough to duplicate what organic minds

do. But if so, there would be a characteristic of protein that is in principle duplicatable in some other substance, even if it was not duplicatable in fact.

(3) Every physical part of a mental system will possess not only those characteristics that are essential to its function in that system, but also other characteristics that are irrelevant to that function. I will refer to the former set of characteristics as *functional*, and the latter as *epiphenomenal*. Strictly speaking I should probably refer to them as *relatively epiphenomenal*. Some recent philosophical discourse has defined "epiphenomenal" to mean absolutely epiphenomenal, that is, irrelevant to every possible causal system, not just to one particular system. Because I believe this meaning of the term is useless, and probably empty, I will use the term "epiphenomenal" to mean relatively epiphenomenal. This is the way it is often used in scientific discourse (see Dennett 1991, p. 402), where it helps to make distinctions similar to the one I am making here. When any biological research goes beyond describing morphology to developing explanations, it must make a distinction between those characteristics that perform functions (like the connections between axons and dendrites in a brain) and those characteristics that are merely epiphenomenal (like its gray color and lumpy shape). The epiphenomenal characteristics will of course have causal properties in other contexts. It is just that these properties will not be in any way responsible for the emergence of mental processes.

(4) A science of mental processes must concern itself with distinguishing between (1) those characteristics of a thinking–feeling creature that perform functions that help constitute mental processes, and (2) those characteristics that are epiphenomenal with respect to mental processes. We will be able to judge what is functionally essential and what is epiphenomenal with regard to mind only if we know the pattern of systematic structure that actual and possible minds share, regardless of what they are made of physically. Note that I am using the words "pattern" and "structure" here in the broadest possible way, so that to say anything at all about why something has a mind would be to articulate a pattern of some sort. Paradoxically, despite the fact that these assumptions require us to see mental processes as something abstract and distinct from any particular

physical characteristics, they are also the only way to make any form of physicalism coherent. If physical matter is not itself mental, and no pattern can be made out of physical matter that can produce mental processes, then mind would be inexplicable. We would then be stuck with some form of dualism, in which mind magically oozes out of organisms like a glowing fluid.

These four principles are the basis of a bare bones functionalism that would be considered trivially true by both functionalists and eliminative materialists. These are the ground rules of the search for that pattern in physical stuff that embodies mind, or is identical with mind, or on which mind supervenes. (We'll deal with the differences between these three descriptions in later chapters.) The disagreements between the functionalists and eliminative materialists arise only because this is a discussion of what future science may look like, and only research can decide between the various possibilities.

The eliminative materialists will admit when pressed that there will probably always be separate sciences of psychology and neuroscience to study the functional and physical characteristics of mind respectively. (Although they sometimes point out that neuroscience *could* eliminate psychology, and that the psychology of the future will probably have even less resemblance to folk psychology than the psychology of the present.) The functionalists will admit that of course one needs to study neuroscience to learn how psychological functions are implemented (although they disagree, with each other and with the eliminative materialists, as to how independent psychology and neuroscience can be). The only real difference between the two camps is who their heroes are, and where they search for scientific facts to bolster their arguments. Eliminative materialists admire the "wet" neurosciences that study actual neurons, and functionalists admire the computer sciences and artificial intelligence. Conflict arises between them when either group presumes that the cognitive science of the future will most resemble their own favorite science of the present. But if things turn out as I believe they will, many of these surface differences will vanish. Future wet cognitive science will have to stop focusing on the cranial region of the nervous system and pay attention to the rest of the organism and the environment. AI will recognize that the multiple realizability of functional categories does not entail the autonomy of an inner language of thought, which means AI will also

have to pay more attention to the whole organism and the environment. What we really have here is a conversation masquerading as an argument.

The main thing that keeps both functionalism and eliminative materialism at loggerheads with each other is that each has embraced a slightly different form of Cartesian materialism, neither of which is essential to the basic program they both share.

Some Cartesian Materialist Presuppositions

When Patricia Churchland says "I am a materialist and hence believe that the mind is the brain" (P. S. Churchland 1986, p. ix), she does not treat this assertion as a position to be defended, but as an uncontroversial given that would be accepted by all factions of the materialist camp. But the fact that most of the eliminative materialists do accept this assumption shows that they are not being completely true to their own principles. As long as they claim that the mind is the brain, they are in fact still identity theorists, and I believe that this alleged identity actually shackles us to certain concepts from folk psychology that could seriously hamper future scientific growth. Eliminating the one-to-one correspondence between mental states and brain states was a step in the right direction, but to be truly consistent they should have also called into question the identity of the mind as a whole with the brain as a whole. I will try to show in this book that careful analysis reveals even the current state of neuroscientific knowledge no longer fully supports this identity, although the presuppositions of both philosophers and scientists have made it very difficult to see this.

A perfect example of this kind of confusion is seen in two essays from the anthology *The Mind–Brain Identity Theory* (Borst 1970). In one of these essays, the classic "Is Consciousness a Brain Process?", U. T. Place recognized that the mind–brain identity claim could not be defended on philosophical grounds alone, and should be considered only as a reasonable scientific hypothesis (p. 42). However, in "Sensations and Brain Processes," J. J. C. Smart dismissed this call for caution by saying that "If the issue is between a brain thesis, or a heart thesis, or a liver thesis, or a kidney thesis, than the issue is purely an empirical one, and the verdict is overwhelmingly in favor of the brain." The verdict is in favor of the brain,

however, only if we assume that the mind must be identical with one particular giblet in the body, as folk anatomy divides it. (Note that all of the alternatives that Smart lists can be found in any Oxford butcher shop.) However, as Patricia Churchland points out "the available theory specifies not only what counts as an explanation, but also the explananda themselves" (P. S. Churchland 1986, p. 398). In other words (my words, not hers), advanced neuroscience will not just give us more information about what the brain does and how it does it. It could also end up eliminating the whole concept of brain, just as easily as it could eliminate any other concept originally derived from folk psychology.

The functionalists have also made a commitment to their own brand of Cartesian materialism, usually unconsciously. Fodor, however, is quite explicit in this commitment when he claims that psychology must accept what he calls "methodological solipsism" (Fodor 1987). What he means by this is that mental states must be studied as an independent system that takes place entirely within a brain, which can be understood without any reference to the outside world. Paul Churchland almost breaks free of this assumption when he points out that even the most radical eliminative materialist must endorse functionalism "construed broadly as the thesis that the essence of our psychological states resides in the abstract causal roles they play in a complex economy of internal states mediating environmental inputs and behavioral outputs" (1989, p. 23). But the grammatical structure of the definition, as well as the fact that he focuses so heavily on brain data in his own work, reveals a commitment to the assumption that the internal states are the only real subject matter, not the environment and the behavior. The functionalist usually sees the system that is receiving inputs and giving outputs as a self-contained system, rather than a dependent pattern that gets its cognitive and biological significance from the context in which it dwells. This gives rise to a myth that is closely related to what Ryle called the "ghost in the machine."

I call this myth, with a similar deliberate abusiveness, the myth of the "machine in the machine." It is the basis for Fodor's "language of thought" theory of mind, and any other theory of mind that holds that all we need to do to understand the mind is to open Skinner's "black box," without worrying about how its contents relate to the organism, environment, and society in which it functions. One of Ryle's biggest mistakes, for which he

has been justly criticized by cognitive philosophy and science, is conflating these two myths. Cognitive science is quite right to claim that the latter is nowhere near as bad a myth as the former. But precisely because it has been accepted for so long, the machine-in-the-machine myth has recently begun to reveal its weaknesses, many of which Ryle accurately foresaw in his original conflated attacks on both myths.

Ryle's Dispositional Psychology

Ryle specifically attacks the machine-in-the-machine myth when he rejects the belief that we can know about minds through "a process of inference analogous to that by which we infer from the seen movements of the railway signals to the unseen manipulations of the levers in the signal box" (Ryle 1949, p. 52). There is obviously nothing ghostly about a signal box. It is every bit as physical as a computer. However, Ryle's rejection of the signal-box analogy makes no distinction between those who believe that minds are brains and those who believe that minds are ghosts. Consequently, much of what he says has no impact on the mind–brain identity theory. Dualists believe that "one person cannot in principle visit another person's mind as he can visit signal boxes" (ibid.), but we can study brains with electrodes, PET scans, and hosts of other technologies that are becoming more sophisticated all of the time. Even if one does not want to describe these methods as "visiting" the brain, we use similar methods to learn about protons and neutrons, even though we will never have knowledge by acquaintance with them. So why should the fact that we can't visit the mind stop us from studying it?

Ryle then says something that could be used as an objection to this reply. We already know a great deal about minds, even though the science of psychology is still in its infancy. So how could we be dependent on some sophisticated theory for this knowledge, the way the physicist is dependent on a sophisticated physical theory? The answer that Paul Churchland gave to this question decades later was based on an insight of Wilfrid Sellars: we know what we know about minds thanks to a *folk*-psychological theory which, like many kinds of folk theories, has respectable predictive power even though it is theoretically confused. Ryle managed to avoid this conclusion by saying that the concepts we think of as being mental are not theoretical, but dispositional, and that it is the

nature of our best dispositional theories to be able to predict by something like conceptual inference. To say that someone is in pain simply means that they have the propensity to perform pain behaviors, that is, wince, cry out, and perform a variety of other actions whose exact boundaries are not delineated, but which everyone knows. This is one of the reasons that Fodor and others referred to Ryle's position as *logical* behaviorism. Fodor was able to come up with a convincing argument why logical behaviorism did not exclude the mind–brain identity theory, and why it could not give any sort of answer that could satisfy science.

He pointed out that many questions have both a causal answer and a logical answer. For example, the question "What makes Wheaties the breakfast of champions?" could be answered by saying that they are full of vitamins and protein. But it could also be answered by saying that a nonnegligible number of champions eat them for breakfast (Fodor 1975, p. 7). Similarly, when we say "Jones is in pain," we mean he's behaving like he's in pain, and our concept doesn't have to be significantly more informative than that to be effective in ordinary discourse. But Fodor points out that although this is fine for common sense, it would never do for scientists to give explanations of this sort. It would be like the police saying that they have recently discovered that the robbery was the work of thieves. This would not be an acceptable answer from a policeman, even if it was fleshed out with further conceptual analysis like: "The tell-tale signs are there: the stolen property, the loss of the moneyed substances, it all points to thieves."[2] Or a more famous example, it would be like saying that opium puts people to sleep because it has dormative powers. Science sets itself the goal of giving a causal story of why Jones is in pain, and for that, Fodor claims, we must look inside the brain.

A Rylean Alternative to Functionalist Cartesian Materialism

Here is where I part company with Fodor, and to some degree rejoin Ryle. For there is no necessary connection between a causal story of mind and the mind–brain identity theory. The basic dogma of Cartesian materialism is that only neural activity in the cranium is functionally essential for the emergence of mind. This implies that all of the behavioral elements that take place in the world, which Ryle considers to be the essential constituents of mind, are actually epiphenomenal with respect to mind,

and consequently a brain in a vat would be conscious even if it never interacted with a body to cause behavior. This might be true, but it is an empirical claim and, as we shall see, one that is perhaps uniquely difficult to prove.

Why should we assume that all human behavior is caused by a machine that lives in our skulls? Many of Ryle's criticisms of this assumption are as valid as ever. Most of what makes a clown's clowning clever takes place in the circus ring, not in the clown's brain. As Andy Clark pointed out again several decades later, when we do math on paper whatever is happening "in our heads" is not sufficient to solve the problem, even though it is necessary (Clark 1997). That is why we need the paper. So why assume that the brain is a closed causal system that creates mind and thought all by itself? Why not say that the mind is dependent on the causal interactions of the brain, the body, and the world?

Ryle was not able to conceive of this possibility, because he wanted to see talk about minds as being reducible to talk about dispositions. Science is no longer willing to accept dispositional "explanations" the way it did in Aristotle's and Molière's time. Science today posits the existence of unseen theoretical entities that are more ontologically fundamental than the variety of dispositions that each one explains. These entities are not unseen because they are very small, a fact that is blurred by the frequent use of atoms as the paradigmatic scientific entities. They are unseen because they are abstractions that enable us to make sense out of higher-level generalities. Theoretical entities are not inside the perceptible entities whose behavior they explain, they are above and beyond them. Gravity is not just the disposition possessed by apples that makes them fall. It has rules of its own that explain the behaviors of apples, planets, and acrobats in ways that are impossible to reduce to discourse about any one of the items that it affects. The mind–brain identity seems inevitable when we see scientific entities as being like atoms. The fundamental assumption of atomism is that we understand things by breaking them down into parts. When we ask the question "What part of the body is the mind?" the best answer to that question may be "the brain." But that is the wrong question if the mind is not part of the body, but rather a pattern that emerges when a living body interacts with a world. In that case a mind would not be any sort of organ. It would be what Dewey called a system of tensions, and what is now called a dynamic system by philosophers like

Tim Van Gelder. We needed gravitational and magnetic fields to go beyond Aristotelian physics to modern physics. Perhaps the thing that is holding psychology back is that it is not yet thinking in terms of "behavioral fields."

Dormative powers are completely ontologically dependent upon sleep and so cannot provide an explanation that meets modern scientific standards. Ryle tried to expand the concept of disposition by saying that "some dispositional words are highly generic or determinable, while others are highly specific or determinate" (Ryle 1949, p. 118). The specific and determinate dispositions are referred to with expressions like "Wheaties-eater" or "cigarette smoker." To say that a Wheaties-eater eats Wheaties is clearly a tautology. But to say that a doctor performs surgery, or that a solicitor drafts wills, is to name only one of the many ways that a doctor can be a doctor or a solicitor can be a solicitor. Ryle, however, is completely silent on how we know which activities of a generic disposition belong to a particular genus, and which don't. Why is it that we know that surgery and writing prescriptions are forms of doctoring, and that tap dancing and water skiing are not? Why is it that if we see a doctor using a brand new surgical technique, we will probably know that he is doctoring even though we have never seen that technique before? The obvious answer, which Ryle occasionally comes close to acknowledging,[3] is that we have a concept of doctor that is more than the sum of the discrete dispositions in which it manifests. Similarly, because we cannot make sense out of human behavior without a theory that posits mental entities that are more than the sum of the individual human actions, we must operate on the assumption that minds are ontologically distinct from those actions.

Ryle's failure to reduce mind to purely dispositional terms shows that even folk psychology is not satisfied with a purely dispositional explication of mind. Infants do learn a folk-Aristotelian dispositional psychology in the (western European) nursery, as Peirce claimed, but they learn a great deal else as well. They learn a causal theory about the mind, which enables them to predict human behavior based on a theory that posits the existence of mental entities like beliefs and desires. Ryle was wrong to think that this causal theory could be reduced without remainder to a list of dispositions. But I believe he was right that this causal theory is not about brains any more than it is about ghosts, despite a few expressions that have trickled down into folk psychology from Hippocratean neuroscience. Any

theory of mind must be a theory about human beings interacting with each other and with their environment. Ryle was also wrong to think that a commonsense theory of mind could be completely independent of neuroscience. Sellars's scientific realism was intended largely as a critique of ordinary language philosophers like Ryle, to remind them that common sense never has this kind of independence. It is always being changed by new scientific discoveries. But Ryle was right to object to the idea that science will ever be able to replace folk psychology with a theory that talks *only* about brains. There is no question that brains are an essential part of the puzzle, but there is also no reason to assume that they are the entire puzzle. If every brain that ever existed did nothing but what brains do in vats (release neurotransmitters, shift blood flow, etc.), then no one would ever have thought that brains had anything to do with minds at all.

In fact, if we take Ryle's famous Oxford University example seriously, we might very well decide that locating the mind in any single organ was a category mistake. Even if we performed rigorous quantitative tests that proved that the administration building controls and directs all of the activity in the other buildings, and that all of the really important classes are given there, this would not prove that Oxford university was identical with the administration building, because buildings and universities are members of different categories. If Oxford decided to rent out a local theater to hold especially large classes, that theater would be part of what was identical with Oxford while the classes were being held there. It would not be identical with Oxford when the town drama society held amateur theatricals there. Similarly, the brain (or the retina or the spinal chord) would be identical with the mind when it performed mental functions, and identical with the body when it performed physical functions (if these two are separable from each other at all). It would be a mistake to argue over whether the theater was part of the drama society or part of the university. It would be a similar mistake to claim that any one part of the body was the mind if the entire body was participating in mental functioning in varying degrees and ways.

When Armstrong put forth his version of the mind–brain identity theory, he paid homage to the Rylean position he was criticizing by saying it was an essential step in a dialectical process.

> . . . classical philosophy tend to think of the mind as an inner arena of some sort. This we may call the Thesis. Behaviorism moved to the opposite extreme: mind was

seen as outward behavior. This is the Antithesis. My proposed Synthesis is that the mind is properly conceived as an inner principle, but a principle that is identified in terms of the outward behavior it is bringing about . . . if we have . . . general scientific grounds for thinking that man is nothing but a physical mechanism, we can go on to argue that the mental states are in fact nothing but physical states of the central nervous system (Borst 1970, p. 75).

There is a tension in this paragraph that captures the essential error of the mind–brain identity theory; an error that was compounded by the practices of the eliminative materialists who followed the identity theorists. For if mental states are identified by the outward behavior they produce, it seems inevitable that they will be ontologically constituted by that outward behavior as well. Thus we cannot make the claim that mental states are nothing but brain states and keep the synthesis described by Armstrong above. On the contrary, this claim produces a return to the thesis, not a synthesis of the thesis and antithesis. The thesis does change: it becomes Cartesian materialism rather than Cartesian dualism. But it does not incorporate the antithesis and thus resolve the conflict, it continues to "think of the mind as an inner arena of some sort." And this Cartesian materialism justifies the view of mind as an inner arena with a non sequitur, for it does not necessarily follow from the claim that people are physical mechanisms that mental states are in fact nothing but physical states of the central nervous system. On the contrary if we accept the synthesis described by Armstrong above, the inner principle and the outward behavior together constitute the mind.

This conclusion is especially unavoidable with those aspects of mind that are closely associated with language. Externalist philosophers of language, such as Putnam and Burge, have argued that meaning cannot be in the head, because language has an intentional relationship to the world of the speaker (i.e., it is "about" the world). If the word "Paris" means what it means because it has a relationship to Paris, surely our thoughts about Paris must have a similar relationship to Paris. And if so, how can our thoughts be nothing but neurological processes confined to a cranium?

Many contemporary philosophers are claiming that they cannot. Jerry Fodor tried to have it both ways by saying our thoughts consisted partly of a narrow content that supervened only on our brains (Fodor 1981) but later rejected this idea (Fodor 1994). Ruth Millikan has created a detailed critique of what she calls "meaning rationalism" (the belief that meanings

exist only in the head), and in her 1993 she makes the point that "I no more carry my complete cognitive systems around with me as I walk from place to place than I carry the U.S. currency system about with me when I walk with a dime in my pocket" (p. 170). Hubert Dreyfus has introduced a whole generation of scientists and analytic philosophers to Heidegger's idea that an essential characteristic of mind is "Being-in-the-world," and consequently that no self is strictly distinct from the world in which it dwells. Andy Clark makes a similar claim in his book *Being There: Putting Brain, Body, and World Together Again*.

But even though these and many other thinkers are willing to locate verbalizable conceptual thought partly in the world, they are usually not willing to make the same step for feelings, sensations, and conscious experience. Clark gives detailed arguments for showing that language and other forms of cognitive activities could not be seen as self-contained languages of thought within the skull. But in a section entitled "Where does the mind stop and the rest of the world begin?" he deliberately refuses to apply the implications of his argument to subjective experience: "I assuredly do not seek to claim that individual consciousness extends outside the head . . . conscious contents supervene on individual brains. . . . Thoughts, considered only as snapshots of our conscious mental activity, are fully explained, I am willing to say, by the current state of the brain" (Clark 1997, pp. 215–17).

In Clark and Chalmers 1998, Clark gets a bit bolder and merely concedes that "some mental states, such as experiences, *may* be determined internally" (italics added). What I am claiming here is that Clark and the other externalist philosophers of mind have not been bold enough. I am not merely repeating the slogan in Putnam 1975 that "Meaning ain't in the head." I am also saying that "Consciousness ain't in the head." Most of the strongest objections against externalism can be best dealt with by completely rejecting the distinction between intentional mental processes and so-called raw feels. All experience is, I claim, completely and irreducibily intentional, and thus gets its meaning from relationships that the living self maintains with the outside world.

I do not mean by this that all experience is really linguistic. This misinterpretation of Sellars has many contemporary defenders, most prominently Dennett and Rorty. (Fodor is not included in this company only because he refuses to say anything at all about consciousness.) But I am

not one of them. I believe that although language and experience are both intentional, they relate to their intentional objects in importantly different ways. To my knowledge, Dewey was the first philosopher to recognize and describe these differences. He showed why an intentional theory of experience would be safe from both the incoherent concept of "raw feel," and the dangerously oversimplified view that language is all there is to mind. What we have learned since his time, both philosophically and scientifically, has made his intentionalist view of experience more relevant than ever.

Is this perhaps only a philosopher's quibble? It is not immediately obvious that a laboratory neuroscientist needs to worry about this question at all. After all, the idea that the mind is the brain is not really that far off. Doesn't modern neuroscience confirm Hippocrates' claim that most of the important mental processing occurs in the skull, even if we have to acknowledge, if pressed, that perhaps someday we may discover that not all of it does? So what's the big deal? Do the confusions in the mind–brain identity theory really lead to any important philosophical or scientific confusions? Not perhaps in the short run, but in the long run, such confusions can lead to crisis and sometimes scientific revolutions.

There is no denying that the mind–brain identity theory works, in a rough and ready sort of way, just as folk psychology works in a rough and ready sort of way. But it may very well be that most mental functions have been found in the skull only because that is where people have been looking for them. As Kuhn has taught us, the paradigm always sets the rules for the puzzles that normal scientists must try to solve. If one of the rules is "Don't run experiments that test for mental functions anywhere but in the brain," the fact that almost no cognitive action has been found anywhere else doesn't really prove that much. Perhaps if laboratory researchers began recognizing that the mind is not simply the same thing as the brain, they might look elsewhere for mental functions, and they might find them. It *could* be a matter of contingent fact that all mental functions are performed exclusively by a single organ or system of organs. But we will never find evidence proving or disproving this claim if we assume it to be true before we begin.

Unless experiments are performed that are expressly designed to falsify the claim that the mind is the brain, we cannot say that this claim has been scientifically established, no matter how natural it may seem to us.

Nor, for that matter, can my suggestion that the mind is distributed elsewhere be any more than a suggestion until experiments are performed that are designed to falsify it. My only claim is that a noncranial mind is a genuine empirical possibility, and not an empty logical possibility of the sort that interests no one but philosophers. The assumption that the mind is the brain will probably always be a useful working hypothesis for certain forms of research. But my hunch is that it may someday be seen to resemble Newtonian physics when compared to Einsteinian physics. In other words, there may be certain kinds of data that can only be accounted for by a whole new theory. If such a theory does become necessary, we will have to concede that the mind–brain identity theory is false in scientifically significant ways: that, strictly speaking, there can be no mind without a brain–body–world nexus. The hope is that serious attention to this possibility will either confirm or refute it. In the next two chapters, I will look at some scientific data that do seem to be pointing in that direction, despite most scientists' lack of interest in the fact. In later chapters, I will claim that many of the most stubborn philosophical paradoxes arise from the unconscious (and arguably unjustified) assumption that science has proven that the mind exists only in the head.