Finally, I attempt to give a balanced presentation of the issues, but I make no pretense of neutrality. It is my view that an author gains nothing by trying to be invisible. Ideas are the property (and properties) of thinking beings, not just disembodied possibilities of belief. I am skeptical of many of the philosophically stronger claims coming out of AI research; I am one who suspects that the nature of the mind may remain mysterious for a long time yet. Nevertheless, my skepticism is not meant to be pessimistic. The sciences of the mind may be young, but the philosophy of mind can still learn much from them. If I can give to the reader a sense of the difficulty of this terrain and the excitement of trying to traverse it, I will consider my efforts successful.

Chapter 2

The Mind-Body Problem

Grasping the Problem

Human persons are creatures who, like all creatures, have a physical, biological nature. Our bodies are an important part of what we take to be essentially human about ourselves. We recognize and understand the capacities and limitations of our bodies, and we even measure the passage of time by our sense of growth and aging.

Already, though, language exerts a kind of centrifugal force upon thinking; it pulls away from complete identification with the body. We say, "I have a body," which implies that whatever I am, it is not quite the same thing as my body, since that which has cannot be exactly the same as what is had. This may appear to be a mere linguistic quirk, but it points to a much larger philosophical issue. That is, it points to the fact that it is very basic to our way of thinking about ourselves that we are not just bodies, not just creatures with a biological nature. Whether or not you share this intuition as a part of your personal philosophical system, there can be no doubt that it is embedded in the worldview of the West. How it got to be so embedded is a question that we shall take up later in this chapter.

Let's take a closer look. Consider the moment of waking up in the morning. Perhaps you have been dreaming, but the quality of your dream experience changes somehow. Some sounds that were in your dream now have the character of being external to your dream, and in fact the dream is now over. Your eyes flicker open to take in a complex of
color and shape, which you understand without a moment’s hesitation to be the real world. You are seeing and hearing, already engaged in interaction with the world. Moreover, your thinking is probably already detached from the specific sensory information being presented to you. You are thinking about whether to get up or lie there for a few more minutes, what to have for breakfast, what to wear, what to do today, and so on.

This detachment of your thinking from your perception—an obvious and constant fact of human experience—is very important. It marks a kind of cleavage of experience into two parallel streams. First, there is the stream of experience that you take to be the record of the world impinging upon you, the world of sight, sound, and feeling. Then there is the stream of experience that we usually refer to as “thinking.” The literary term “stream of consciousness” is somewhat misleading, since it misses the cleavage that we have just noticed. Thinking and perception are both things that take place in the stream of consciousness, yet they are separate somehow.

Both streams are streams of experience; that much is very clear. Together, they appear to constitute the total of a person’s experience while awake. But these two streams are also clearly different. We need to consider what is different about them, and why we take them to be two streams instead of one in the first place.

First, there is the matter of control. You have very little control over the way things sound or look. To be sure, you can close your eyes or put your fingers in your ears, giving you an indirect kind of control, but you have almost no direct control over the exact character of your perceptions. On the other hand, your thinking does appear to be very much, though not absolutely, under your conscious control. You can decide what to think about, to some extent, and what not to think about. Of course, it is important not to overstate the extent of your sovereignty over your thinking. Try not thinking about monkeys for the next minute. Consider

1. This presupposes a fairly liberal understanding of perception and thinking, of course. There may be good reason later to narrow our understanding of these terms, but for now the point should be clear enough.

the number of times you have had a fragment of a song or an advertising jingle “stuck” in your thinking for the better part of a day.

Still, even the limited control that you have over your thinking is far more extensive than the control that you have over perception. This difference in control is therefore one intuitive basis for dividing the stream of experience into two streams. Now consider another aspect of this division. When you are with other people, you can talk about your shared perception of the world. That is, your perceptions are confirmed by other perceptions, in a certain way. If you see a tree, you can say to the person with you, “Do you see that tree?” If you hear the person say “Yes,” then that is another perception of yours, which confirms the previous visual perception that you just had. Why do you take the second perception to be a confirmation of the first? It is because you believe that what you perceive to be another person is in fact another person like yourself who has the same kind of experiences. Your question, and that person’s answer, mark what is taken to be the public character of the world. The world is out there for all to see, and that is why we expect to agree on such things, for the most part. This is sometimes referred to as “consensus reality.”

When it comes to your own thoughts, the other stream of experience, things are quite different. You do not expect that some other person can confirm that you are thinking about lunch at this moment, or that the wine tastes good to you, or that you have a headache. When you ask, “Is it warm in here or is it me?” you are appealing to this same distinction between the two streams of experience. In short, your thoughts are the scenery of your “inner life,” the private domain of your own mind.

It is interesting to note that we often use words of containment to talk about the mind. We speak of what is “in” the mind, the “inner life” as opposed to the “external world.” We may say that these are just metaphors, figures of speech. Perhaps if so, they have a powerful hold upon us, because it is hard not to think of the mind as a kind of sealed-off space, to which the rest of the world is somehow external. If these labels seem arbitrary, try for a moment to think of your mind as external, while the things around you are internal. Perhaps for a moment it is possible to achieve this perspective, but it requires (for me, at least) a powerful
burst of imagination. Yet it seems so natural to divide the world the other way, into an inner mind and an external reality. In everyday speech, when we say that something is “all in your mind,” we mean that it isn’t real, not a part of consensus reality.

As I say, this division seems natural, even inevitable. But it is precisely this division that gives rise to the mind-body problem. After all, what we have been calling the external stream of experience is more simply understood as our awareness of objects in the external world. This includes not only the book in your hands but your hands themselves and the rest of your body. Your body is, after all, from one perspective just one more bit of the furniture of the external world. You can see it and others can see it; it is part of consensus reality. Even your brain, considered for the moment simply as part of your body is an object in the external world. It is not generally visible to the naked eye, but it can be made visible. During brain surgery a section of your skull may be removed, allowing you (with the aid of mirrors) and everyone else to see it. Your mind, in the sense of an inner stream of experiences, is something else again. Your intuitive sense of the matter is likely to be that your mind is not just that brain that you see in the mirrors, that the brain is external to your mind. How can both of these somehow be you? What is it about this particular body that makes it your body and not someone else’s? How does your mind come to enjoy such an intimate and unique relation to your body? What is that relation?

Continuing to work from the level of intuition, some further observations can be made. For one thing, the body isn’t just there; it clearly has effects on the mind. I have already mentioned perception as giving rise to one stream of experience. This process involves the direct participation of the body. It is light striking the retina of your eye, or sound waves upon your eardrum, that initiate the sequence of events that somehow cause (or become) your perceptual experiences. This is an example of events in your body causing events in your mind.

At other times you may think about what to do, then decide what to do, then do it. Here the sequence begins with events in your mind—thinking—which bring about events in your body—your voluntary behavior or action.

2. There are unusual individuals who can exert a certain amount of voluntary control over the activity of their heart, but this only shows that the details of which physical events can be affected by which mental events is subject to considerable variation. Loss of muscle function by paralysis shows the same thing.
Classical Theories

Platonic Dualism

Plato is usually cited as the first Western philosopher to offer a theory of mind and body, even though he was not working from a sense of the mind-body problem as sketched above. In the dialogue called *Phaedo*, Plato (narrating the words of Socrates, his teacher) attempted to give an account of the possibility of the immortality of the soul. What Plato (and Socrates) meant by the soul was not quite the same as what that word has come to mean in religious contexts, nor is it the same as what we have been calling "mind" in the discussion above, but it is related to both of these.

Plato apparently believed that one of the soul's most important attributes is its capacity for abstract intellectual thought, while one of its least important attributes is its capacity for perception. In fact, Plato believed that what we have been calling the "external world"—the visible world of concrete objects and processes—is only part of reality, and by no means the most important part. In addition to the visible world there is, thought Plato, an "intelligible world" of abstract ideas and relationships that exist independently of the mind. What makes this notion philosophically interesting is the fact that Plato offered a fairly sophisticated argument in defense of it. It is worth taking the time to review that argument.

Imagine two metal rods, apparently of equal length. I say "apparently" because you will probably concede that if you were able to subject the rods to a sufficiently precise microscopic analysis, you would find that they are in fact not equal in length. If you think about it, you might also concede that no two rods are ever exactly equal in length.

That's fine, but Plato went on to ask how, then, we have any concept at all of what "exactly equal" might mean, if we never actually encounter it in the visible world. How could we know that a given pair of rods is closer to "exactly equal" than another pair, if exact equality is something that we have never experienced? Plato believed that since we do have a concept of exact equality it follows that we do have some experience of it, even though that experience could not be based in the visible world.

It is as an abstraction, an object of intellectual appraisal, that we understand what exact equality is; the objects of the visible world may approach it, but they never quite reach it.

So, in order to account for our understanding of certain facts about the visible world, such as the fact that no two objects are ever exactly equal in length, Plato had to suppose that there is another world, populated by objects of a different sort: abstract objects. He called these objects "forms." While the senses perceive the objects of the visible world, it is the mind (or soul) that "perceives" or apprehends the objects of the intelligible world.

Plato took the argument further. Although the objects of the visible world are subject to change, decay, and extinction, the objects of the intelligible world are eternal and unchanging. The concept of exact equality, for example, is what it is, regardless of the variation in visible objects. Since the human mind has a kind of access to this realm of eternal and unchanging ideas, it must itself be in that world. And since decay and death are phenomena of the visible world only, it follows that the part of a person that is of the intelligible world would not suffer these processes. Conclusion: the mind/soul is immortal, while the body is mortal.

There is no space to consider fully the details and problems of this argument. What is important about it is that it divides all of reality into two categories of being, with the mind in one category and the body in another. This move is called dualism, and this particular flavor of dualism is called, not surprisingly, Platonic dualism.

Cartesian Dualism

As I mentioned, Plato's dualism was not specifically a response to the mind-body problem; it was a response to other philosophical questions, such as the possibility of immortality and man's true nature. In fact, it would not be far wrong to say that our sense of mind and body posing a problem is a result of Plato's dualism, since it directs our attention to these issues in a focused way.

In the seventeenth century, the French polymath René Descartes
CHAPTER 2. THE MIND-BODY PROBLEM

articulated the mind-body problem in a more formal way, and it is
Descartes's formulation that has come to serve as the basis for modern
examinations of the problem. Descartes was also the first to appreciate
the possibility of automating much (but not all) of human behavior, and
is therefore one of the earliest figures in the philosophy of artificial
intelligence.

Like Plato, Descartes was not specifically interested in the mind-body
problem; he was working out a larger plan. The seventeenth
century was a period of intense political and intellectual ferment. The
Protestant Reformation had already occurred, and as a result the Roman
Catholic church could no longer be viewed as the only possible source of
legitimate opinions on important questions. Important scientific discov-
eries were being made that challenged the traditional theology-
ized view of man and his place in the universe. The
"Copernican revolution," according to which mankind's place is not in
the center of the universe, was not just a revolution in astronomy; it was
a revolution in thought, dislodging a centuries-old picture of the world.

Descartes, like many intellectuals of this period, was concerned
about the breaking up of the unified picture of the world that had been
in place for so long. He himself made important contributions in science
and mathematics and was as aware as anyone of the way in which
human knowledge, while growing, was no longer held together in a
coherent system. This need to see things as elements of a coherent
system is perhaps the deepest expression of the philosophical impulse.

What Descartes set out to do, then, was to find foundations for knowl-
edge. That is, he hoped that he could show that all human knowledge,
no matter how diverse and apparently unrelated, was in fact related by
resting on a common set of foundations.

Foundational knowledge, if it could be shown to exist, would have to
be the most basic possible knowledge and, since a structure is only as
strong as its foundations, it would have to be as certain as possible. To
this end, he devised what has come to be called the "method of doubt." The
application of the method is fairly simple: For each thing that you
might claim to know, ask whether it could conceivably, under any
imaginable circumstances, be subject to doubt. If the answer is yes, then
put that bit of alleged knowledge aside, as if it were false; it cannot serve
as a foundation of all knowledge. Apply this method systematically and
see what, if anything, does not get put aside. If anything survives the
method of doubt, it will be an appropriate foundation of knowledge, since
it is immune to doubt.

Consider the knowledge that you get by using your senses. You would
claim to know, for example, that you have a book in your hands right now,
and that knowledge-claim is based on a sensory experience you are
having. Could this knowledge-claim be subject to doubt under any
imaginable circumstances? Note that this is not the same as asking
whether you do in fact doubt it at this very moment. Rather, the question
is: Could you doubt it?

Of course you could. It is not hard to think of cases in which you are
completely misled by your senses. To give a somewhat far-fetched exam-
ple, you might be hallucinating this book. Or you could be dreaming that
you are reading a book. It doesn't matter that you don't for a moment
believe that you are hallucinating or dreaming; what matters is that you
concede that you could be. If so, you must reject sensory knowledge as
foundational.

Perhaps, like Plato, you would turn to the realm of the pure intellect.
You would claim to know that \(2 + 2 = 4\). Could you be mistaken in this
sort of knowledge? It's hard to see how you could, but Descartes went so
far as to suppose that there is some demon who deceives us whenever we
set out to do any mathematical thinking. Again, the point is not that
there is such a demon, but that there could be. If knowledge of the pure
intellect, such as mathematical knowledge, is susceptible to error in this
way, then it cannot serve as a foundation of all knowledge.

It seems that if Descartes is going to allow such wild possibilities as
hallucination and demonic intervention to play a role, then nothing is
truly immune to doubt. So it seems, but is it precisely this point that
Descartes introduces his most famous proposition: I think, therefore I
am. That is, whatever else it may be possible for me to doubt, I cannot
doubt my own existence. To do so would be to be thinking about my own
existence, and if I am, then there must be someone doing the thinking. The very act of
doubting my own existence proves it beyond any possible doubt, argued Descartes.

I begin, then, with my own existence as a genuinely foundational piece of knowledge. Descartes went further, to reflect upon what sort of thing it is of whose existence I am now assured. When I conclude that I think, therefore I am, what is the very minimum that I am entitled to assert about the nature of this "I"? Since the existence of "I" is established by the fact of thinking, it would follow that the one thing I know about "I" is that it is a thinking thing. "I am a thing which thinks," wrote Descartes, dubbing this entity with its Latin name, res cogitans. I cannot doubt my own existence as a thinking thing. But I can doubt my own existence as a physical entity extended in space—res extensa. The only basis I have for concluding that my body exists is sensory, and sensory knowledge has already been shown by the method of doubt to be uncertain.

Now follows the key move, in relation to the mind-body problem. If my body is capable of being doubted but my mind is not capable of being doubted, then my mind has at least one property that my body lacks: immunity to doubt. Furthermore, this immunity to doubt was supposed by Descartes to be an essential property of the mind. That is, it is a property without which it could not even be a mind (as opposed to an "accidental" property, which a thing might or might not have). If mind and body do not have all the same essential properties, then they cannot be one and the same thing; they must be different things.

Once again, we end up with dualism. The world contains two (at least) kinds of things: thinking things and things extended in space, res cogitans and res extensa. Since mind is not extended in space, it cannot be said to have any specific location, but since thinking does have duration, it follows that mind is extended in time.

This form of dualism is called, appropriately, "Cartesian dualism." It is the form of dualism that has been most influential in modern discussions of the mind. It would be fair to say that Cartesian dualism has become the basis of the "common sense" understanding of the mind.

Descartes visited the Royal Gardens at Versailles and saw there the world-famous exhibit of automata. These were mechanical animal and human figures that turned and moved in what seemed at the time to be lifelike ways, driven by the hydraulic power of water pumped from the fountains. Descartes was impressed. He was so impressed, in fact, that he decided that the physical processes of the human and animal bodies must operate according to similar principles. The body, on this view, is a complex mechanical apparatus. He even described a hydraulic model of behavior. It was known that the brain contained fluid-filled cavities and that a clear fluid would drip from cut nerves. It was reasonable to propose that the muscles move the limbs by hydraulic power, with the reservoirs of hydraulic fluid in the brain.

Still, Descartes would not go so far as to suppose that all human experience could be reduced to mechanical-hydraulic principles. He asserted, for example, that no automaton of any complexity could learn to use language, that for that task a mind is required.

Although it may not seem so now, it was a bold stroke of imagination for Descartes to conceive of a biological organism as an automaton, and even bolder to conceive of the person as such an automaton linked somehow to an immaterial mind. It is only with this picture in place that modern discussions of the philosophy of mind and artificial intelligence make sense.

Recent Approaches

One of the most striking features of the worldview that emerged from the breakdown of the authority of the Church and the emergence of science was mechanism. Newton's laws of motion purported to describe not just the behavior of this or that object or system, but of any physical object or system at all. These laws involved a careful bookkeeping of forces, masses, momenta, and energy. Just as the behavior of a clock is driven by the energy stored in its wound mainspring, the behavior of any physical system should not exceed, in its expenditure of energy, the sum of energy available to it. Since clocks were complicated but well-understood devices, the metaphor of world-as-clockwork had a great appeal.

In more contemporary language, we say that energy is "conserved"
in a closed system. A closed system is one in which no energy is allowed to enter or leave, so the conservation law says simply that, under those conditions, no energy will appear or disappear; the net amount of energy in the system will remain the same.

Now think of the human brain, and the entire human organism, as such a system. Whatever happens there must, according to the law of conservation of energy, be accounted for by energy available to the system. In this case, that energy is in the form of chemical energy derived from foods eaten. It is not at all clear what the role of an immaterial mind could be in this system. If, as Descartes supposed, the brain is the controller of a complex hydraulic system, then that system must be driven by physical energy. How could a mind, which has no mass or size or even location, exert any influence on a mechanical system? How could the "actions" of a mind make any difference at all?

This is the "problem of interaction," and Descartes himself was aware of it. His "solution" was to suppose that there is some physical region where the interaction takes place, an organ devoted to just that purpose. His candidate was the pineal gland, a small almond-shaped organ located on the center axis of the brain, three inches or so in back of the point between the eyebrows. Here, he thought, the mind could move the "animal spirits"—a hypothetical energy-like fluid—which would in turn make the body work.

This is no solution at all, of course. For one thing, the problem of interaction is hardly solved by saying that the mind only interacts directly with a part of the brain, since it was not the size of the brain that was the problem. The problem is to show how a massless, sizeless, immaterial mind could have any effect at all on any physical system, large or small. If the mind is adding energy to the system somehow, then it ought to be possible in principle to detect that empirically.

In short, at the same time that Descartes was creating the commonsense picture of the mind, he was laying the groundwork for the eventual opposition to his theory. So successful was his image of the body as a mechanical device that it was not long before people began to wonder if the more problematic part of his model, the immaterial mind, was necessary at all. In 1749 the French physician and philosopher Julien

Offrey de la Mettrie published L'Homme Machine (Man the Machine), in which he promulgated a thoroughly mechanistic philosophy of the person, a view offensive to many of his contemporaries.

Materialism

The principal alternative to dualism is materialism. According to materialism, there are not two categories of things in the world but only one: material, or physical, things. Everything that happens in the universe involves physical objects, forces, and processes, and nothing stands outside of that totality of physical interactions. For all physical phenomena, physical causes must be sought and can, in principle at least, be found.

Materialism is the metaphysics of modern science, for the most part. That is, most working scientists are convinced that there is nothing in the world that is not subject to and ultimately explained by the laws of physics. It is important to recognize that although Descartes worked near the beginning of the scientific age, his dualism was almost a desperate last stand against the tide of materialism. Although dualism may remain basic to the common-sense picture of the mind, a wide gulf has opened between that common-sense picture and the scientific picture.

The twentieth-century philosopher Gilbert Ryle, in his book The Concept of Mind, lampooned the Cartesian conception of mind as a "ghost in the machine." What he meant was that the materialistic conception of the person is complete in itself; it is unnecessary to suppose that there is some additional entity called a "mind" to make the thing work. In fact, according to Ryle, to suppose that the mind must be some sort of "thing" or entity is to commit a "category mistake." A category mistake is the result of grouping something in a category with other things that are logically dissimilar. To use an example of Ryle's, if you visit a university and are shown the classroom, dorms, administration buildings, and so forth, and then ask, "But what about the university," you are making a category mistake. The university is not something in addition to its grounds, buildings, students, faculty, and the like; these
things and people just are the university. On this view, the mind is not a substance or entity, immaterial or otherwise, that is somehow above or beyond the organic brain; it is an aspect or property of the physical organism.

It is important to recognize that in general—materialism claims something about what minds are. In fact, materialism is not a single well-defined claim at all; it comprises a number of variations, with important differences. Minimally, however, materialism says that nothing beyond what we understand to be the physical universe is involved in minds and mental processes.

Since materialism is actually a class of theories, it will be useful to sketch some of the more important variations.

The Identity Theory

According to this theory, minds and brains are identical or, alternatively, mental states are brain states. This is much stronger than claiming that mental states are correlated with brain states; a dualist would very likely agree with the latter claim. The identity theorist is claiming that mental states and brain states are one and the same. This claim implies that the terms "mental state" and "brain state" refer to exactly the same thing, just as "Peking" and "Beijing" refer to the same city.

Although the identity theory seems admirably clear and simple, there are some difficulties associated with it. To understand those difficulties, it will first be necessary to take a closer look at Descartes’s position on dualism. Recall that Descartes’s argument for dualism pivoted on the claim that mind and body cannot be one and the same if their essential properties are not the same. This claim is based on a principle that has come to be called the "indiscernibility of identities." If A is identical to B, then all properties of A must be properties of B, and vice versa. Descartes argued that since the body is subject to doubt and the mind isn’t, the two are not indiscernible, and hence not identical.

Descartes’s argument has subsequently been criticized for the way in which it depends upon the beliefs of the person involved. To make this criticism clear, consider the following example: Is it possible that you could be certain that the Prime Minister of Britain is alive but doubt that John Major is alive? Yes, it is possible if you don’t know that Mr. Major is the Prime Minister. Would it follow, from the fact that you can doubt that Mr. Major is alive, that he is not the Prime Minister? Obviously it wouldn’t.

By similar reasoning, the fact that the body’s existence—but not the mind’s—can be doubted doesn’t prove anything about whether the mind and body are identical. Does this invalidate the principle of indiscernibility of identities? No, but it shows that the principle needs to be clarified a bit. For A and B to be identical, what is required is that they share all properties in common; we now know that that is too strong. Rather, A and B have to have all properties of a certain sort in common. The trick is to say just what sort of properties count in identity claims and what sort do not.

Intuitively, what we want are only those properties that do not depend upon what somebody believes about a thing, since what one might or might not manage to doubt clearly depends upon what one believes. The fact that somebody might doubt that Mr. Major is the Prime Minister doesn’t tell us anything about whether he is or isn’t the Prime Minister. Let’s call those properties that a thing has that do not depend upon what anybody believes about it its objective properties. Now we can restate the principle of indiscernibility of identities as follows: A and B are identical if and only if every objective property of A is an objective property of B, and vice versa. If we then claim that being subject to doubt is not an objective property, Descartes’s argument for dualism fails. This doesn’t mean that dualism is false, of course, but only that Descartes’s particular argument doesn’t support it.

Getting back to the identity theory, the revised version of the principle of indiscernibility of identities can be applied to the claim that

---

3. In technical philosophy, the matter is discussed not in terms of properties but in terms of predicates that truly apply. The particular kind of predicates that correspond to what we are calling "objective properties" are "extensional predicates." I mention this only because this technical language is fairly common in the literature on the subject, even though introducing it here would make matters unduly complicated.
mental states are identical with brain states. If even one objective property of mental states can be found that brain states lack, or vice versa, then the identity theory is certainly false.

One property that has been much discussed in this connection is privileged access. That is, I seem to have a kind of access to my own mental states that no one else can have. I know my mental states from the inside, as it were. In fact, when I first introduced the mind-body problem at the beginning of this chapter, I did so in terms of a division of experience into two streams. One stream is part of “consensus reality” while the other is not. I claimed as a basic intuition that I have a special access to this latter stream. That I have this access appears to be an objective fact about my mental states. My brain states, however, are publicly accessible, as was stated earlier; they are there for anybody to observe.

So, if I really do have privileged access to my mental states (and not my brain states) and if privileged access really is an objective property, then the identity theory is just wrong. Predictably, both suppositions have come under attack. Some philosophers have denied that there is any such thing as privileged access, while others have disputed whether any such thing could properly be called “objective.”

You may wonder how anyone could challenge the claim that one has privileged access to one’s own mind. The identity theorist may say that there is no reason in principle why someone’s mental states could not be revealed by a microscopic inspection of his brain (assuming that we knew what to look for). If that is true, then there is nothing that that person could think that could not be discovered by someone else.

Still, you might argue, we couldn’t come to know these things in the same way—that is, by introspection—that he knows them. True, but the identity theorist will respond that that is beside the point. If someone can determine (reliably) what another is thinking, then regardless of how it is done, the other’s access is not privileged.

Another issue connected with the identity theory is the clarification of the actual terms of the identity equation. So far, I have been using the terms “mental state” and “brain state” as if it were intuitively obvious what these are. But there is an added distinction that is very relevant to the meaning of these terms: the distinction between type and token.

A type is simply a category of objects, whereas a token is a particular object. The word “apple,” for example, refers to a type of fruit. When we say something like “this apple,” we are usually referring to a token of that type—a particular apple. You can eat apple-tokens, but you can’t eat apple-types (even though you can eat apples of different types).

The distinction is clear enough as it refers to apples, but is often forgotten when the subject is mental and brain states. Consider the mental state of believing that it will rain within an hour. It is possible that one might have such a belief many times, so it is reasonable to refer to the belief that it will rain soon as a belief-type. On the other hand, it makes equal sense to think of it as a belief-token, if you think of the belief as a particular event that occurs at a particular moment in time. Clearly, belief-tokens are unique objects that occur once, while belief-types may be “tokened” any number of times. Beliefs are mental states,4 but the same type-token distinction applies to brain states.

It is one thing to say that every mental state-token is identical to some brain state-token. It is quite another thing to say that every mental state-type is identical to some brain state-type. Since tokens by definition never recur, the token-token identity theory asserts no more than that whenever a person is in some mental state, she is in some brain state or other, and that that mental state just is that brain state. Many versions of dualism would allow that every mental state is correlated with some brain state; it is the identity claim that the dualist feels goes too far. What the token-token identity theory does not allow is disembodied mental states, or minds without brains.

The type-type identity theory is much stronger. According to it, whenever a person is in a mental state of a particular type, she must also be in a brain state of some corresponding physical type, and to change mental state-types is to change brain state-types. As in the token-token version of the identity theory, the claim of correlation is strengthened by

---

4. Or events. The difference between a state and an event is that an event is a sequence of one or more states in time.
an identity claim: Mental states of a particular type just are brain states of a particular type. It should therefore be possible in principle to make lawlike claims of the form, "If A is in a mental state of type M, then there must also be some identifiable type of brain state B1 that A is in." Furthermore, if it makes sense to say that two individuals, A and B, are ever in the same mental state, such as the state of believing that it will rain soon, then it must follow that their brains are in states of the same physical type.

But this is all very implausible. For one thing, what evidence is there that my brain is in the same state whenever I believe that it will rain soon? I could conceivably be in any number of brain states and have that belief. Worse, it is very unlikely that two people are ever in precisely the same brain state, if we take a brain state to be a specific pattern of activity and excitation of neurons. Perhaps the notion of "same state" needs to be relaxed somewhat. The problem is understanding just how to relax it. What we cannot do is say that two brain states are "the same" if they support (or embody, or are-identical to) the same mental state. We cannot say this because it begs the very question at issue.

Carrying these objections further, it would seem to follow from either version of the identity theory that anything without a brain must also be without mental states. But what is a brain? It is a type of biological organ with a certain molecular and chemical structure. We can easily imagine being visited by intelligent extraterrestrials with organs that do the same kinds of things that our brains do but whose molecular and chemical structure is quite different. It seems that the identity theory would require that we deny that these beings could have any of the same mental states that we have. When the E.T. takes out an umbrella, we would have to say, "It's as if he believes it's about to rain."

You might want to change the definition of "brain" to make it looser on chemical and biological details but more explicit on functional ones. Perhaps a brain is "anything that controls a body." Such a move widens the field so much that it no longer deserves to be called a "mind-brain identity theory."

It is also clear that it would be impossible in principle for computers to have mental states, since computers are not brains. If artificial intelligence involves any artificial mental states—and many supporters of artificial intelligence would insist that that is precisely what it involves—then according to the identity theory artificial intelligence is impossible. For the reasons given, such an approach seems unduly restrictive.

This is as far as I plan to go with this sketch of the identity theory. Many of the issues involved—especially the matter of privileged access—will surface again in later chapters. Since the identity theory is not the only version of materialism to be considered, it is important to consider some others.

**Behaviorism**

Ryle's "ghost in the machine" was an expression that perfectly captured the disdain that many scientists and philosophers had for "mentalist"—the attempt to place mental states and objects into a protected category. Dualism is perhaps the clearest expression of materialism, but it is not the only expression of it.

To understand behaviorism as a reaction against mentalism, it is necessary to understand something of the philosophical and scientific movement called "logical positivism." This movement was—and in some ways continues to be—tremendously influential in shaping the scientific philosophy of the twentieth century.

Logical positivism was an attempt to bring philosophy into line with the tough-minded empirical methods that had proven so successful in the sciences, especially physics. It was widely accepted that that success was due to the rigorous application of the "scientific method," which involved, among other things, testing all hypotheses by observation and experimentation.

In the period between the two world wars, a group of scientists and philosophers had informal meetings in Vienna to discuss how philosophy and science might be somehow merged into a single intellectual endeavor. This group, which came to be known as the "Vienna Circle," comprised such individuals as A. J. Ayer, Moritz Schlick, and Rudolf Carnap. They recognized that before a proposition could reasonably be
said to be true or false, it must be cognitively meaningful. “Cognitive” meaning is the sort of meaning that is relevant to the truth conditions of a proposition. Other, “noncognitive” varieties of meaning are possible, on this view, but not relevant to truth or falsity. If a proposition is not meaningful, it could not have truth conditions and therefore could not really be a proposition. The first problem, then, was to say precisely what it is for a proposition to be meaningful. If this could be accomplished, meaningful propositions could be distinguished from meaningless pseudo-propositions by applying an objective criterion. Science and scientific philosophy would, of course, restrict their attention to meaningful propositions.

First, it was noted that certain propositions are formally true. That is, their truth is a consequence of their meaning within some formal system. The concept of a formal system is the subject of the next chapter. For the moment, it is sufficient to say that a formal system comprises a set of symbols and precisely defined rules for manipulating them. The meaning of the symbols in the formal system is deliberately assigned, in an interpretation. The various computational techniques that we call “arithmetic” belong to a formal system. Thus, the proposition \(7 + 5 = 12\) is perfectly meaningful and true in that formal system, or formally true. In that same formal system, \(7 + 5 = 13\) is meaningful but false, and \(7 + 5\) is not even meaningful, hence neither true nor false.

The logical positivists proposed the Verification Principle for those propositions that do not derive their meaning from a formal system. Here is a simple formulation of the Verification Principle:

A nonformal proposition is meaningful if and only if it is in principle verifiable by observation.

To “verify” a proposition in this context may be taken to mean “determine whether it is true or false.” The words “in principle” are very important, since there are many propositions that are impossible in practice to verify, such as the proposition “There is intelligent life in the Andromeda galaxy.” But as long as it is possible in principle to verify it by observation, the proposition is meaningful. The kind of observation that the positivists had in mind was public observation; that is, the verifying observation should be accessible to any observer, in principle.

You may be thinking that there was no need to make a special category for the formal propositions, since \(7 + 5 = 12\) is apparently verifiable by observing simple acts of counting. Although there are good reasons for doubting this, a clearer example would be the following: The sum of every pair of odd numbers is an even number. Since this proposition applies to infinitely many cases, there is no question of verifying it by observation.

Now consider the proposition “Colorless green ideas sleep furiously.” Is it meaningful? The Verification Principle says that for it to be meaningful, there must be circumstances under which one could observe it to be true or false. But it is hard to conceive of any such observation in this case. What would it be like to observe a colorless green idea furiously sleeping, or falling to? It seems safe to say that this proposition is not verifiable by any conceivable observation. So, according to the Verification Principle, it is quite meaningless and neither true nor false.

The application of the Verification Principle to these sample cases is clear enough. The logical positivists made a much more extensive and ruthless use of it. One philosopher, C. L. Stevenson, applied it to statements of moral judgment, such as “It is wrong to tell a lie.” Since it does not seem possible to observe the wrongness of lying, Stevenson concluded that it is in fact not a meaningful proposition at all in the scientific or “cognitive” sense, but rather an expression of the utterer’s emotional attitude toward lying, disguised as a proposition. So the statement has only noncognitive meaning. This positivistic theory about the meaning (or lack thereof) of moral language was called “emotivism.” Similar moves were made for aesthetic judgments. If beauty is in the eye of the beholder, then its presence or absence is not publicly observable.

A domain of inquiry in which a rigorous application of the Verification Principle was apparently overdue was psychology. At first, J. B. 6. Not everything that you can count will follow the rules. Seven quarts of water plus five quarts of pure alcohol does not equal twelve quarts of solution.
CHAPTER 2. THE MIND-BODY PROBLEM

Watson, the founder of psychological behaviorism, offered it as a methodological reform. How, he asked, can a science study such things as ideas, beliefs, and desires if these things are not, in principle, observable? Psychology is supposed to be the science of the mind, but how can there be such a science if it is impossible to formulate meaningful propositions about the objects of the mind?

Consider belief. Beliefs are mental objects, most would agree. I may see you leaving your house in the morning carrying an umbrella and wearing a raincoat, and I may hear you say to me, “I think it’s going to rain today.” I would certainly conclude that you believe that it will rain today, but do I observe that belief? No. What I observe is what you wear, what you carry, and what you say. I infer the belief from these observations, but I do not observe it.

Science should deal only in cognitively meaningful propositions, which must be verifiable by observation. Beliefs, ideas, and other so-called “mental objects” are not observable. Behavior is observable. The conclusion is clear: The science of psychology should study behavior. This argument became the foundation of psychological behaviorism.

In the form in which I have just stated it, psychological behaviorism neither affirms nor denies the existence of the mind and mental objects; it merely places the domain of scientific study elsewhere. And in fact it brought about a great methodological revolution in psychology. A kind of behavioristic orthodoxy grew up, with journals, psychology faculties, and graduate students controlled by it. At the same time, behaviorism was undergoing a kind of purification in the hands of philosophers.

Why stop with methodological recommendations? Why not go all the way and claim that propositions that purport to refer to mental objects either are meaningless or, at best, do not mean what people have usually taken them to mean?

Positivism wanted to do for psychology what it had done for ethics. The problem was that while emotivism said that statements in the form of moral judgments were really expressions of the speaker’s emotions, it was not clear what statements about beliefs and ideas were really expressing. The best solution was to propose that all mental terms really refer to behavior or dispositions to behave, and that the mentalistic language evolved simply as a shorthand. So, when I say that so-and-so believes it will rain today, what I really mean is that so-and-so is engaged in a certain kind of behavior (umbrella carrying, raincoat wearing, and the like) and nothing more.

If behavior is all that we really mean when we use mentalistic language, then you might wonder why it took so long for us to figure this out. The philosophical behaviorists used the analogy of sunrise and sunset. When people use the words “sunrise” and “sunset,” all they are ever really referring to is the position and motion of the earth relative to the sun. They are not, and never have been, referring to the motion of the sun around the earth, because there is no such motion. The fact that for most of recorded history people didn’t know what “sunrise” and “sunset” properly referred to doesn’t alter these facts. Likewise, when we say “believe” we are referring to behavior, whether we acknowledge it or not.

This version of behaviorism has come to be called philosophical behaviorism. From this vantage point, the whole mind-body problem is dismissed as a pseudo-problem, the result of a confusion about the real meaning of the language that we use. The word “mind” is just a convenient way of referring to a complex array of behaviors and dispositions, and the various other mental terms refer to subsections of that array and the relations between them. To assert that the mind is not the body is no more than to assert that the behavior is not the same as the thing that is doing the behaving, a harmless assertion. To suppose that the mind could exist independently of the body would be to suppose that there could be behavior in the absence of anything to do the behaving, a preposterous supposition. End of problem.

Philosophical behaviorism is classified under materialism because it is an essentially materialistic philosophy that gave rise to logical positivism in the first place, and ultimately to behaviorism. The material world is paradigmatically observable, and propositions about it are paradigmatically verifiable. It was the success of physics, after all, that provided the fuel for the positivistic movement. To a certain extent,
logical positivism can be understood as the attempt to put the rest of human understanding in a position to imitate and partake of that success.

I have deliberately presented behaviorism as if its rise over the intellectual horizon were smooth and unimpeded. As you might expect, this was not at all the case. At every step of the way, logical positivism was challenged, as was behaviorism, from within and without. The Verification Principle itself was harshly criticized and subjected to endless amendments. While psychological behaviorism had smooth sailing for a while, since working psychologists were excited by the objective research program to which it pointed, philosophical behaviorism was embattled from the start. Still, philosophical behaviorism has left its mark on the mind-body problem. Indeed, it returns in a new guise in many discussions of artificial intelligence, as we shall see in the chapter that deals with the Turing Test.

Philosophical behaviorism faces a number of difficulties. For one thing, if a proposition is verifiable in principle by observation, the number of observations that count toward verifying it ought to be finite. Returning to the example of the person who has the belief that it will rain today, it is obvious that the statement "Joe believes it will rain today" cannot be translated into a statement about any single behavior. The number of behaviors that could be taken to "translate" this statement is quite large, arguably infinite. To have the belief that it will rain today is to be disposed to carry an umbrella or to wear a raincoat or to say "I think it will rain today" or to say "I believe it will rain today" or to wear galoshes or ... . There seems to be no way to restrict the range of possible behaviors that are the meaning of this simple belief. The theory collapses under its own weight.

Functionalism

It has seemed to many philosophers that while philosophical behaviorism and the identity theory face serious difficulties, there is something right about both of them. The identity theory seems to be on the right track, if we could only get a less restrictive account of what a brain state—or a brain, for that matter—is. Behaviorism at least has the right sort of skepticism about mental objects, if we could only get a better account of what those objects actually are.

The theory that attempts to preserve what seems right about behaviorism without getting locked into the restrictions of the identity theory is functionalism. According to functionalism, mental states are just functional states of complex systems. To make sense of this simple claim, it is necessary to say in detail what a functional state is.

Let’s begin with a simple example. Suppose that you and I have radio sets tuned to the same station. Your equipment comprises a modern solid-state tuner and amplifier, using integrated circuit technology. I have an antique radio that uses vacuum tubes and coils. There is a large difference in the quality of the sound produced by the two sets, but they are nevertheless producing the "same" music. Using a more general term, we may say that they are producing the same "output." They are also making use of the same input: the fluctuations in the surrounding electromagnetic field called "radio waves."

Since the input and output are the same, it follows that in some sense what both devices are doing with the input is the same: They are decoding it to produce music. But how they are doing it is rather different. So in one sense the two devices are doing the same thing; in another sense they are doing very different things. To understand these two senses, we need to introduce the idea of "levels of description" of a system.

This is actually a fairly complicated and subtle idea, but a simplified account of it should suffice to make the point about functionalism. Consider the two radio devices mentioned above. If we confine ourselves to a physical description of what they are and what they are doing, it is clear that they don’t have much in common. They are made of different materials, and those materials are configured as parts in very different ways, and the physical interactions between those parts is also very different. At a higher level, however, we might just look at various parts of the system and describe them in terms of what they do. This part detects the radio signal; this part amplifies it; this part works the speakers. At that level of description, there is a correspondence between
what the two devices are doing. And we have already granted that at the
level of producing output they are doing just the same thing (though not
equally well). We may call the higher level of description the “functional
level,” since it has to do with how a thing functions, rather than what it
is made of.

We may further say that any device that detects radio waves and
decodes them to produce audible music is functionally equivalent to any
other device that does the same thing. Two systems are in functionally
equivalent states if they can accept the same input and “process” it
somehow to produce the same output.

Now imagine a third device, somewhat unusual in construction. This
one is capable of detecting and decoding radio waves, but it is not able to
send an amplified signal to the speakers. It does have a built-in cassette
player, however, that can drive the speakers and produce audible music.
Suppose that the music recorded on the cassette were exactly the same
music that happened to be being played by the other two radio devices.
So, while those two devices are playing broadcast music, this third device
is playing a cassette recording of the very same music that is being
broadcast. Is this third device functionally equivalent to the other two?

It is not. Even though it detects the same radio waves and is playing
the same music as the other two, it is clear that there is not a causal link
between the detection of the waves and the production of that music—it
just happens that the inputs and outputs of all three devices are, for the
time being, the same. What is different is the relation between input and
output; the character of that input is not the cause of the specific
character of the output.

Now consider how these concepts relate to the mind-body problem.
According to functionalism, to be in a particular mental state is just for
your brain (or central nervous system) to be in a particular functional
state. If another brain is in that functional state as well, then it too is in
that mental state. Furthermore—and here is the improvement on the
identity theory—if something other than a brain, such as a digital
computer, is in the same functional state, then it must also be said to be
in the same mental state.

Functionalism is indifferent to what sort of system is producing the
functional states, as long as it is sufficiently complex to produce them at
all. This is an important part of its appeal from the standpoint of
developing a philosophy of mind that can accommodate the possibility of
artificial intelligence (and extraterrestrial intelligence).

Note that this is not simply a disguised version of behaviorism. Men-
tal states are not, according to functionalism, just output or dispositions
to produce certain output. Rather, to be in a given mental state is to be
processing input in a certain way to produce certain output. Consider the
range of states that we call “mental” — beliefs, ideas, hopes, fears, desires,
and so forth. Functionalism says that when a being—or system—is in
any of these states, it must also be in some particular functional state,
and that there is a “type-type” correspondence between mental and func-
tional states. Functional states are identified by the causal relations that
they bear to each other and to inputs and outputs. Perhaps you see an
apple, which reminds you of the color of your red sweater, which in turn
reminds you that you must pick that sweater up at the dry cleaner’s,
which gets you to pick up the telephone to check to see if the sweater is
ready. This is a long causally linked chain of mental states: a perception,
a memory, another memory, a decision, and a volition. If functionalism
is true, each of these states is really just a functional state of your brain,
causally linked in an analogous way. If we could get some other system to
go through the same sequence of functional states, then we would have
to attribute the same mental states to it.

It may be useful to attempt a slightly more formal definition of
functionally equivalent systems.

A state, S1, of system A is functionally equivalent to state S2 of
system B if and only if A and B will reliably produce the same
subsequent state and/or output, given the same input.

This causal isomorphism between mental and functional states im-
plies that the two kinds of states will be individuated in the same

7. A doughnut, for example, just doesn’t have the right sort of complexity to
produce the functional states that your brain can produce.

8. That, of course, is just one way of describing the sequence.
way. Where one mental state ends and another one begins must also be where one functional state ends and another one begins. This requirement is, as we shall see later, a point on which functionalism has been widely criticized. Without it, however, there is no chance of revealing any sort of regular mapping between mental and functional states. Without such a mapping, it would be impossible to say that anyone who is in mental state $M$ must also be in functional state $F$. And if we can’t say that much, it is hard to see that functionalism has any very interesting content at all. The condition that $A$ and $B$ will reliably produce the same subsequent state or output is there to eliminate systems that just happen to do so.

At the moment, however, functionalism remains the favorite theory of the mind among artificial-intelligence advocates. Rather than enumerating criticisms here, therefore, I shall take them up as they are relevant to issues in later sections. It is also important to keep in mind that, strictly speaking, functionalism is neutral on the issue of materialism versus dualism. The functionalist has no official position on what a system must be made of to have functional states. If a Cartesian immaterial mind can have them, then that is what matters. This neutrality is somewhat disingenuous, of course; it is simply a way of sidestepping the ontological issues involved. It will come as no surprise to learn that most functionalists are staunch materialists who believe that only physical systems can have functional—and therefore mental—states.