

---

## The Problem of Phenomenal Consciousness

Science has pushed man farther and farther from the center of the universe. We once thought our planet occupied that center; it doesn't. We once thought that our history was more or less the history of the world; it isn't. We once thought that we were created as the crown and guardian of creation; we weren't. As far as science is concerned, people are just a strange kind of animal that arrived fairly late on the scene. When you look at the details of how they work, you discover that, like other life forms, people's bodies are little chemical machines. Enzymes slide over DNA molecules, proteins are produced, various chemical reactions are catalyzed. Molecules on the surfaces of membranes react to substances they come into contact with by fitting to them and changing shape, which causes chemical signals to alter the usual flow of events, so that the machine's behavior can change as circumstances change.

Traditionally there was one big gap in this picture: the human mind. The mind was supposed to be a nonphysical entity, exempt from the laws that govern the stars, the earth, and the molecules that compose us. What if this gap closes? What if it turns out that we're machines all the way through?

This possibility may seem too implausible or repugnant to contemplate. Nonetheless, it looms on the horizon. For some of us, it seems like the most likely possibility. The purpose of this essay is to increase the plausibility of the hypothesis that we are machines and to elaborate some of its consequences. It may seem that a wiser or more moral strategy would be to avoid thinking about such a weird and inhuman hypothesis. I can't agree. If we are indeed physical systems, then I get no comfort from the fact that most people don't know it and that I can occasionally forget it.

## Mind as Self-Fulfilling Description

I will be arguing that people have minds because they, or their brains, are biological computers. The biological variety of computer differs in many ways from the kinds of computers engineers build, but the differences are superficial. When evolution created animals that could benefit from performing complex computations, it thereby increased the likelihood that some way of performing them would be found. The way that emerged used the materials at hand, the cells of the brain. But the same computations could have been performed using different materials, including silicon. It may sound odd to describe what brains do as computation, but, as we shall see, when one looks at the behavior of neurons in detail, it is hard to avoid the conclusion that their purpose is to compute things. Of course, the fact that some neurons appear to compute things does not rule out that those same neurons might do something else as well, maybe something more important; and there are many more neurons whose purpose has not yet been fathomed.

Even if it turns out that the brain is a computer, pure and simple, an explanation of mind will not follow as some kind of obvious corollary. We see computers around us all the time, none of which has a mind. Brains appear to make contact with a different dimension. Even very simple animals seem to be conscious of their surroundings, at least to the extent of feeling pleasure and pain, and when we look into the eyes of complex animals such as our fellow mammals, we see depths of soul. In humans the mind has reached its earthly apogee, where it can aspire to intelligence, morality, and creativity.

So if minds are produced by computers, we will have to explain how. Several different mechanisms have been proposed, not all of them plausible. One is that they might “excrete” mind in some mysterious way, as the brain is said to do. This is hardly an explanation, but it has the virtue of putting brains and computers in the same unintelligible boat. A variant of this idea is that mind is “emergent” from complex systems, in the way that wetness is “emergent” from the properties of hydrogen and oxygen atoms when mixed in great numbers to make water.

I think we can be more specific about the way in which computers can have minds. Computers manipulate information, and some of this

information has a “causative” rather than a purely “descriptive” character. That is, some of the information a computer manipulates is about entities that exist because of the manipulation itself. I have in mind entities such as the windows one sees on the screens of most computers nowadays. The windows exist because the computer behaves in a way consistent with their existing. When you click “in” a window, the resulting events occur because the computer determines where to display the mouse-guided cursor and determines which window that screen location belongs to. It makes these determinations by running algorithms that consult blocks of stored data that describe what the windows are supposed to look like. These blocks of data, called *data structures*, describe the windows in the same way that the data structures at IRS Central describe *you*. But there is a difference. You don’t exist *because of* the IRS’s data structures, but that’s exactly the situation the window is in. The window exists because of the behavior of the computer, which is guided by the very data structures that describe it. The data structures denote something that exists because of the data structure denoting it: the data structure is a wish that fulfills itself, or, less poetically, a description of an object that brings the object into being. Such a novel and strange phenomenon ought to have interesting consequences. As I shall explain, the mind is one of them.

An intelligent computer, biological or otherwise, must make and use models of its world. In a way this is the whole purpose of intelligence, to explain what has happened and to predict what will happen. One of the entities the system must have models of is itself, simply because the system is the most ubiquitous feature of its own environment. At what we are pleased to call “lower” evolutionary levels, the model can consist of simple properties that the organism assigns to the parts of itself it can sense. The visual system of a snake must classify the snake’s tail as “not prey.” It can do this by combining proprioceptive and visual information about where its tail is and how it’s moving. Different parts of its sensory field can then be labeled “grass,” “sky,” “possibly prey,” “possible predator,” and “tail.” The label signals the appropriateness of some behaviors and the inappropriateness of others. The snake can glide over its tail, but it mustn’t eat it.

The self-models of humans are much more complex. We have to cope with many more ways that our behavior can affect what we perceive. In

fact, there are long intervals when everything we perceive involves us. In social settings, much of what we observe is how other humans react to what *we* are doing or saying. Even when one person is alone in a jungle, she may still find herself explaining the appearance of things partly in terms of her own observational stance. A person who did not have beliefs about herself would appear to be autistic or insane. We can confidently predict that if we meet an intelligent race on another planet they will have to have complex models of themselves, too, although we can't say so easily what those models will look like.

I will make two claims about self-models that may seem unlikely at first, but become obvious once understood:

1. Everything you think you know about yourself derives from your self-model.
2. A self-model does not have to be true to be useful.

The first is almost a tautology, although it seems to contradict a traditional intuition, going back to Descartes, that we know the contents of our minds “immediately,” without having to infer them from “sense data” as we do for other objects of perception. There really isn't a contradiction, but the idea of the self-model makes the tradition evaporate. When I say that “I” know the contents of “my” mind, who am I talking about? An entity about whom I have a large and somewhat coherent set of beliefs, that is, the entity described by the self-model. So if you believe you have free will, it's because the self-model says that. If you believe you have immediate and indubitable knowledge of all the sensory events your mind undergoes, *that's* owing to the conclusions of the self-model. If your beliefs include “I am more than just my body,” and even “I don't have a self-model,” it's because it says those things in your self-model. As Thomas Metzinger (1995*b*) puts it, “since we are beings who almost constantly fail to recognize our mental models as models, our phenomenal space is characterized by an all-embracing naive realism, which we are incapable of transcending in standard situations.”

You might suppose that a self-model would tend to be accurate, other things being equal, for the same reason that each of our beliefs is likely to be true: there's not much point in having beliefs if they're false. This supposition makes sense up to a point, but in the case of the self-model we

run into a peculiar indeterminacy. For most objects of belief, the object exists and has properties regardless of what anyone believes. We can picture the beliefs adjusting to fit the object, with the quality of the belief depending on how good the fit is (Searle 1983). But in the case of the self, this picture doesn't necessarily apply. A person without a self-model would not be a fully functioning person, or, stated otherwise, *the self does not exist prior to being modeled*. Under these circumstances, the truth of a belief about the self is not determined purely by how well it fits the facts; some of the facts derive from what beliefs there are. Suppose that members of one species have belief *P* about themselves, and that this enables them to survive better than members of another species with belief *Q* about themselves. Eventually everyone will believe *P*, regardless of how true it is. However, beliefs of the self-fulfilling sort alluded to above will actually *become true* because everyone believes them. As Nietzsche observed, "The falseness of a judgment is . . . not necessarily an objection to a judgment . . . . The question is to what extent it is life-promoting . . . , species-preserving . . ." (Nietzsche 1886, pp. 202–203). For example, a belief in free will is very close (as close as one can get) to actually having free will, just as having a description of a window inside a computer is (almost) all that is required to have a window on the computer's screen.

I will need to flesh this picture out considerably to make it plausible. I suspect that many people will find it absurd or even meaningless. For one thing, it seems to overlook the huge differences between the brain and a computer. It also requires us to believe that the abilities of the human mind are ultimately based on the sort of mundane activity that computers engage in. Drawing windows on a screen is trivial compared to writing symphonies, or even to carrying on a conversation. It is not likely that computers will be able to do either in the near future. I will have to argue that eventually they will be able to do such things.

### **Dualism and Its Discontents**

The issues surrounding the relation between computation and mind are becoming relevant because of the complete failure of *dualism* as an explanation of human consciousness. Dualism is the doctrine that people's minds are formed of nonphysical substances that are associated with their

bodies and guide their bodies, but that are not part of their bodies and are not subject to the same physical laws as their bodies. This idea has been widely accepted since the time of Descartes, and is often credited to him, but only because he stated it so clearly; I think it is what anyone would come to believe if they did a few experiments. Suppose I ring a bell in your presence, and then play a recording of the 1812 Overture for you. You are supposed to raise your hand when you hear the sound of that bell. How do you know when you hear that sound? Introspectively, it seems that, though you don't actually hear a bell ringing, you can summon a "mental image" of it that has the same tonal quality as the bell and compare it at the crucial moment to the sounds of the church bells near the end of the overture. (You can summon it earlier, too, if not as vividly, and note its absence from the music.) Now the question is, where do mental sounds (or visual images, or memories of smells) reside? No one supposes that there are tiny bell sounds in your head when you remember the sound of a bell. The sounds are only "in your mind." Wherever this is, it doesn't seem to be in your brain.

Once you get this picture of the relation between mind and brain, it seems to account for many things. I've focused on remembering the sound of a bell, but it also seems to account for perceiving the sound as a bell sound in the first place. The bell rings, but I also experience it ringing. Either event could occur without the other. (The bell could ring when I'm not present; I could hallucinate the ringing of a bell.) So the experience is not the same as the ring. In fact, the experience of the ring is really closer than the physical ringing to what I mean by the word or concept "ring." Physics teaches us all sorts of things about metal, air, and vibration, but the experience of a ringing doesn't ever seem to emerge from the physics. We might once have thought that the ringing occurs when the bell is struck, but we now know that it occurs in our minds after the vibrations from the bell reach our minds. As philosophers say, vibration is a *primary quality* whereas ringing is a *secondary quality*.

Philosophers use the word *qualia* to describe the "ringyness" of the experience of a bell, the redness of the experience of red, the embarrassingness of an experience of embarrassment, and so forth. Qualia are important for two reasons. First, they seem to be crucially involved in all perceptual events. We can tell red things from green things because one evokes a red

quale and the other a green one. Without that distinction we assume we couldn't tell them apart, and indeed color-blind people don't distinguish the quale of red from the quale of green. Second, qualia seem utterly unphysical. Introspectively they seem to exist on a different plane from the objects that evoke them, but they also seem to fill a functional role that physical entities just could not fill. Suppose that perceiving or remembering a bell sound *did* cause little rings in your head. Wouldn't that be pointless? Wouldn't we still need a further perceptual process to classify the miniature events in our heads as ringings of bells, scents of ripe apples, or embarrassing scenes?

So far I have focused on perception, but we get equally strong intuitions when we look at thought and action. It seems introspectively as if we act after reasoning, deciding, and willing. These processes differ from physical processes in crucial respects. Physical processes are governed by causal laws, whereas minds have *reasons* for what they do. A causal law enables one to infer, from the state of a system in one region of space-time, the states at other regions, or at least a probability distribution over those states. The "state" of a system is defined as the values of certain numerical variables, such as position, velocity, mass, charge, heat, pressure, and so forth—primary qualities. We often focus for philosophical purposes on the case of knowing a complete description of a system at a particular time and inferring the states at later times, but this is just one of many possible inference patterns. All of them, however, involve the inference of a description of the physical state of the system at one point in space and time from a description of its state at other points. By contrast, the reason for the action of a person might be to *avoid* a certain state. A soldier might fall to the ground to avoid getting shot. People are not immune to physical laws; a soldier who gets shot falls for the same reason a rock does. But people seem to transcend them.

This idea of physical laws is relatively new, dating from the seventeenth century. Before that, no one would have noticed a rigid distinction between the way physical systems work and the way minds work because everyone assumed that the physical world was permeated by mental phenomena. But as the universe came to seem mechanical, the striking differences between the way it works and the way our minds work became more obvious. Descartes was the first to draw a line around the mind and

put all mental phenomena inside that boundary, all physical phenomena outside it.

Nowhere is the contrast between cause and reason more obvious than in the phenomenon of free will. When you have to make a decision about what to do, you take it for granted that you have a real choice to make among alternative actions. You base your choice on what you expect to happen given each action. The choice can be difficult if you are not sure what you want, or if there is a conflict between different choice criteria. When the conflict is between principle and gain, it can be quite painful. But you never feel in conflict in the same way with the principle of causality, and that makes it hard to believe that it is a physical brain making the decision. Surely if the decision-making process were just another link in a chain of physical events it would feel different. In that case the outcome would be entirely governed by physical laws, and it would simply happen. It is hard to imagine what that would feel like, but two scenarios come to mind: either you would not feel free at all, or occasionally you would choose one course of action and then find yourself, coerced by physics, carrying out a different one. Neither scenario obtains: we often feel free to choose, and we do choose, and then go on from there.

Arguments like these make dualism look like a very safe bet, and for hundreds of years it was taken for granted by almost everyone. Even those who found it doubtful often doubted the *materialist* side of the inequality, and conjectured that mind was actually more pervasive than it appears. It is only in the last century (the twentieth) that evidence has swung the other way. It now seems that mere matter is more potent than we thought possible. There are two main strands of inquiry that have brought us to this point. One is the burgeoning field of neuroscience, which has given us greater and greater knowledge of what brains actually do. The other is the field of computer science, which has taught what machines can do. The two converge in the field of *cognitive science*, which studies computational models of brains and minds.

Neither of these new sciences has solved the problems it studies, or even posed them in a way that everyone agrees with. Nonetheless, they have progressed to the point of demonstrating that the dualist picture is seriously flawed. Neuroscience shows that brains apparently don't connect with minds; computer science has shown that perception and choice

apparently don't require minds. They also point to a new vision of how brains work in which the brain is thought of as a kind of computer.

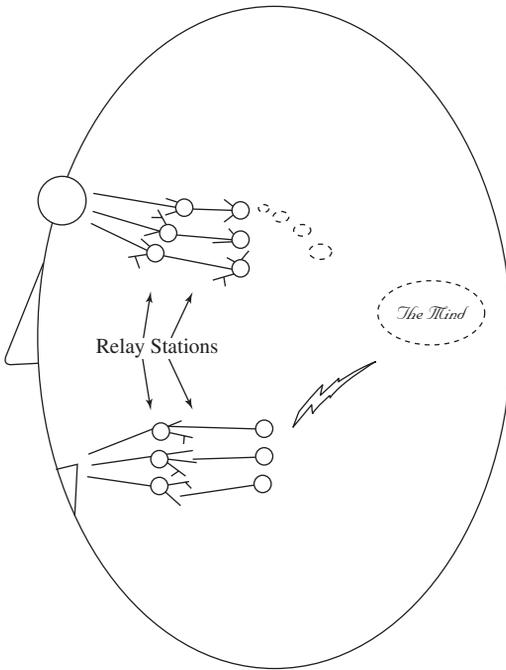
Let's look at these trends in more detail, starting with the brain. The brain contains a large number ( $10^{11}$ ) of cells called *neurons* that apparently do all its work.<sup>1</sup> A neuron, like other cells, maintains different concentrations of chemicals on each side of the membrane that surrounds it. Because many of these chemicals are electrically charged *ions*, the result is a voltage difference across the membrane. The voltage inside the cell is about 60 millivolts below the voltage outside. When stimulated in the right way, the membrane can become *depolarized*, that is, lose its voltage difference by opening up pores in the membrane and allowing ions to flow across. In fact, the voltage briefly swings the opposite direction, so that the inside voltage become 40 millivolts above the outside voltage. When that happens, neighboring areas of the membrane depolarize as well. This causes the next area to depolarize, and so forth, so that a wave of depolarization passes along the membrane. Parts of the cell are elongated (sometimes for many centimeters), and the wave can travel along such an elongated branch until it reaches the end. Behind the wave, the cell expends energy to pump ions back across the membrane and reestablish the voltage difference.

When the depolarization wave reaches the end of a branch, it can cause a new wave to be propagated to a neighboring cell. That's because the branches of neurons often end by touching the branches of neighboring neurons. Actually, they don't quite touch; there is a gap of about one billionth of a meter (Churchland 1986). The point where two neurons come into near contact is called a *synapse*. When a depolarization wave hits a synapse, it causes chemicals called *neurotransmitters* to be emitted, which cross the gap to the next neuron and stimulate its membrane. In the simplest case one may visualize the gap as a relay station: the signal jumps the gap and continues down the axon of the next neuron. When a neuron starts a depolarization wave, it is said to *fire*. Many neurons have one long branch called the *axon* that transmits signals, and several shorter ones called *dendrites* that receive them. The axon of one neuron will make contact at several points on the dendrites of the next neuron. (A neuron may have more than one axon, and an axon may make contact on the dendrites of more than one neuron.) A depolarization

wave travels at a speed between a few hundred centimeters per second and a hundred meters per second, depending on exactly how the axon is configured.

Nowadays we take it for granted that the reason neurons fire is to convey information. That's because we're familiar with the transmission of information in physical forms that are remote from the forms they take when they are first captured or ultimately used. It doesn't strike us as odd that sound waves, disturbances in air pressure, are encoded as little bumps on CDs or electrical impulses in wires. Two hundred years ago this idea would not have been so obvious, and someone looking at the operation of the brain might have been quite puzzled by the depolarizations traveling through neural membranes. Those who take dualism seriously might demand proof that the signals were actually conveying information. Fortunately, it's not hard to find proof. First, we need to show that as the situation is varied in one place in the brain, the behavior of neurons elsewhere in the brain varies accordingly. This phenomenon has indeed been demonstrated over and over. Light is received by the retina, and neurons in the visual system become active; sound is received by the ear, and neurons elsewhere in the brain respond. Of course, it is not enough to show variation. We must also show that there is a *code* of some kind, so that a piece of information is represented by a consistent pattern of neural behavior that is different from the pattern for other pieces of information. I'll talk about that shortly.

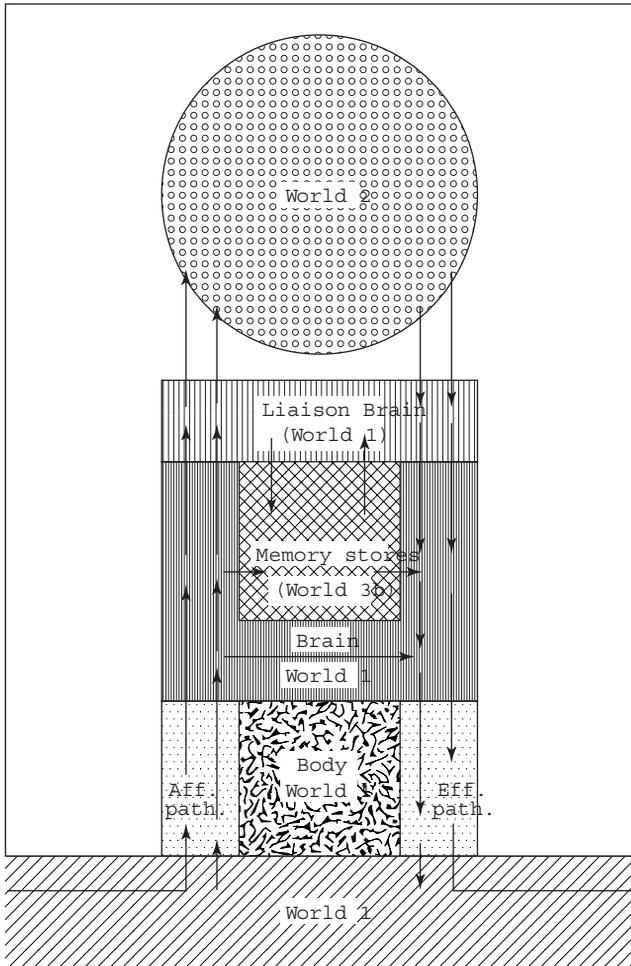
If neural impulses were on their way to a nonphysical mind, one would expect to see neurons transmitting data as faithfully as possible, preserving the content intact until it reached the place where true perception began, and where qualia arose. I've sketched this possibility in figure 1.1. At the point where the interface to the mind appears, one might see neurons with inputs and no outputs. On the "other side" of the mind, one would see neurons with outputs but no inputs, which react to the decisions of the mind by sending impulses to the muscles relevant to the action the mind has decided on. The gap between the last layer of input neurons and the first layer of output neurons might not be so blatant. There might be no spatial gap at all, just a "causality gap," where the behavior of the output neurons could not be entirely explained by the behavior of the input neurons, but would also depend in some way on mental events.



**Figure 1.1**  
The naïve dualist picture

Lest figure 1.1 be thought of as a straw man, in figure 1.2 I have reproduced a figure endorsed by Sir John Eccles, one of the few unabashed dualists to be found among twentieth-century neurophysiologists (Eccles 1970, figure 36, detail). He divides the world into the material domain (“World 1”), the mental domain (“World 2”), and the cultural domain (“World 3”), which I have omitted from the figure. The brain is mostly in World 1, but it makes contact with World 2 through a part called the “liaison brain.” The liaison brain is where Eccles supposes the causality gap lies.

Unfortunately for this dualist model, the behavior of neurons doesn’t fit it. For one thing, there are few places at which data are simply transmitted. Usually a neuron fires after a complex series of transmissions are received from the neurons whose axons connect to it. The signals coming out of a group of neurons are not copies of the signals coming in. They



**Figure 1.2**  
 From Eccles 1970, p. 167, figure 36

are, however, a *function* of the signals coming in. That is, if there is a nonphysical “extra” ingredient influencing the output of the neurons, its effects must be very slight. As far as we can tell, any given input always results in essentially the same output.

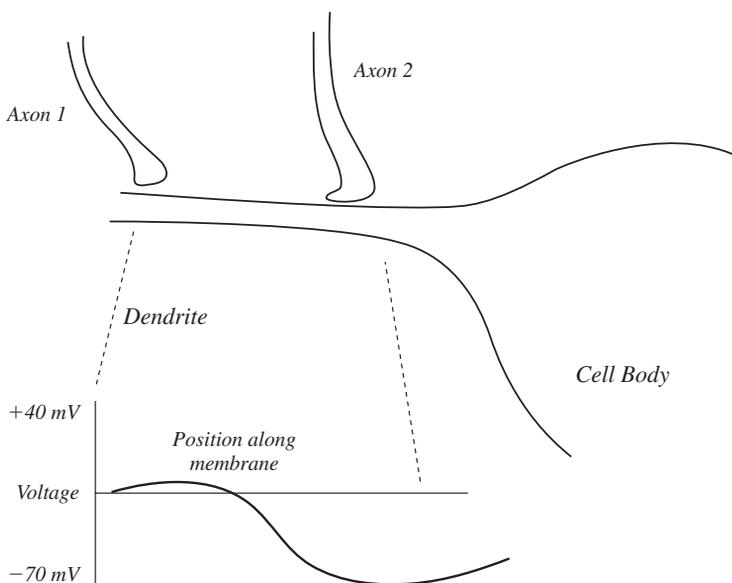
We have to be careful here about exactly how much to claim. Neurons’ behavior changes over time, as they must if they are to be able to learn.

However, such changes are reflected in changes of synapses' sensitivity to inputs, so that the output becomes a function of the input plus the state of the synapses, and their state changes as a function of its input and previous state. More subtly, we must deal with the fact that no system ever behaves exactly the same way twice. There is always a tiny variation in the output. Is this a place where mental effects can slip in?

The answer will be no if the variations are irrelevant to the information carried by the neuronal signals. The only way to judge what's relevant is to understand better how neurons encode information. So far, we understand it only partially. In many cases, the information encoded is the average rate at which the neuron fires. For instance, when light strikes a light-sensitive cell in the retina of the eye, the cell responds by firing several times in succession.<sup>2</sup> It behaves for all the world as if the brightness of the light is encoded by the number of depolarization waves that are transmitted per time unit. The size of each depolarization is irrelevant. Indeed, they are all of about the same size. The membrane either depolarizes or it doesn't. Each depolarization, called a *spike* because of its appearance when charted against time, is rapid and complete. What changes is the number of spikes that occur. The greater the stimulation, the faster the firing rate. Some cells can be negatively stimulated, or *inhibited*, in which case they fire more slowly than their normal rate.

Firing rates are not the only coding scheme in the brain. They could not be, because before a neuron fires its dendrites are combining information from several axons using as a medium only the smoothly varying voltages across postsynaptic membranes. In figure 1.3, the axon on the left has fired and caused the membrane voltage to increase, from its normal value of  $-70$  mV to more like  $-40$  mV. However, the axon on the right is connected to this cell by an *inhibitory* synapse, so that as it fires it "hyperpolarizes" the membrane, driving it even more negative than usual. The net effect depends on the geometry, chemistry, and input firing rates, and is not well understood.

Given that any physical quantity is capable in principle of serving as a code for a piece of information, it might seem impossible to determine if a given physical setup is properly described as actually encoding anything. Fortunately, there is another principle we can appeal to. Suppose someone proposes a certain code for some module of the brain. Then we can look



**Figure 1.3**  
Graded action potentials along dendritic membrane

at the data represented by the inputs to that module under the proposal, look at the data encoded by the outputs, and ask whether the output is an *interesting* function of the input, in the sense that one could see why an organism would want to compute that function. If the answer is yes, then that is evidence that the code is real. Such codes are now found routinely. For example, in the visual system of the brain there are arrays of cells whose inputs represent brightness values at all points of the visual field and whose outputs represent the degree to which there is a vertical brightness edge at each point. Such an edge is defined as a brightness profile that changes value sharply along a horizontal line, from light to dark or vice versa. Other arrays of cells are sensitive to edges with other orientations.

Finding edge detectors like this is exciting because there are independent theories of how information might be extracted from the visual field that suggest that finding edges will often be useful. (For example, if the sun goes behind a cloud, all the brightnesses become smaller, but many of the edges stay in the same place.) But what I want to call attention to is

that the edges are being found *before* the signals “reach the mind.” Edges are not something perceived qualitatively, or at least not exclusively. Here we find edges being found in a *computational* sense that is essentially independent of mind.

At this point we have been led to notice the importance of the second major intellectual strand in the story, namely, the science of computation. We usually use the phrase “computer science” to refer to it, but that doesn’t mean it’s about laptops and mainframes. It’s about physical embodiments of computational processes, wherever we find them, and it appears that one place we find them is in groups of neurons.

Let’s turn our attention from neurons for a second and think in terms of artificial systems. Suppose we build an artificial ear. It takes sound waves, analyzes them for different frequencies, and prints out the word “bell” if the frequency analysis matches the profile for a bell, and “not a bell” otherwise. I don’t mean to suggest that this would be easy to do; in fact, it’s quite difficult to produce an artificial ear that could discriminate as finely as a person’s. What I want to call attention to is that in performing the discrimination the artificial ear would not actually experience ringing or the absence of ringing; it would not experience anything. If you doubt that, let’s suppose that I can open it up and show you exactly where the wires go, and exactly how the software is written. Here a set of tuning forks vibrate sympathetically to different frequencies; here an analog-to-digital converter converts the amplitude of each vibration to a numerical quantity; there a computer program matches the profile of amplitudes to a stored set of profiles. There’s no experience anywhere, nor would we expect any.

Hence it should give us pause if the structures in the brain work in similar ways. And in fact they do. The ear contains a spiraling tube, the cochlea, different parts of whose membrane vibrate in resonance with different frequencies. These vibrations cause receptor cells to send trains of spikes of different frequencies, the frequency of a spike train encoding the amplitude of a particular sound frequency. The overuse of the word “frequency” here is confusing, but also illuminating. One physical quantity, the rate at which a neuron fires, is being used to encode a completely different quantity, the magnitude of a certain frequency in the spectrum of a sound, just as voltages are used in digital computers to represent entities

that have nothing to do with voltages. The representations feed further computations in each case, leading to the extraction of richer lodes of information, such as the phonemes and words hidden in a stream of sounds. What they apparently don't lead to is any experience of anything.

The problem is that every time we find information being extracted computationally, that's one less job for the mind to do. If we find that color discriminations can be done by nonminds, that is, by sensors and computers, and we find the brain doing it the way a nonmind would do it, where does the mind come in? The traditional dualist presupposition was that the brain does almost nothing except prepare raw sensory data to be passed to the mind. That's not the case, so we have to find some other way for the mind to be present.

One possibility is that the mind is still lurking there, it just requires a bit more preprocessing than we used to think. In other words, it receives not raw brightness data, but data already helpfully analyzed into colors, edges, texture analysis, matchings of the images from the two eyes, and so forth. What the mind does is *experience* all these things. Another possibility is that experience inheres in the brain as a whole; in addition to performing computation, the collection of molecules has a quite different function, namely, to provide qualia for some of the discriminations made by the computations.

The question that now arises is what role these experiences play in determining behavior. Dualism maintains that they are crucial. If we trace out the events happening in neurons, we will (at least in principle) find places where what happens cannot be explained purely in terms of physics and chemistry. That is, we would find a gap where a physical event  $P_1$  caused a nonphysical experience  $N$ , which then caused another physical event  $P_2$ . For example, consider the behavior of a wine taster. She sips a bit of 1968 Chateau Lafitte Rothschild, rolls it around her tongue, and pronounces "magnifique." Dualism predicts that if we were to open up her brain and take a peek, we will see neurons producing spike trains that represent various features of the wine. These spike trains apparently would go nowhere, because their sole function is to interface with the intangible dualist mind. Other neurons would produce impulses sent to the mouth and vocal cords to utter the word "magnifique," but these impulses would apparently not be a function of anything. As discussed

above, the gap might be much smaller than depicted in figures 1.1 and 1.2, but it would have to be there, and the link across it would be nonphysical.

It is possible that we will encounter such a linkage in the brain, but almost no one expects to. Eccles (1973, p. 216) proposes that his “liaison brain” is located in the left hemisphere, because the speech center of most people is located in the left hemisphere. Needless to say, despite intense research activity, no such linkage has appeared. Of course, if it existed it would be very hard to find. The network of neurons in the brain is an intricate tangle, and there are large sections as yet unexplored. Almost all experimentation on brains is done on nonhuman animals. Probing a person’s brain is allowed only if the probing has no bad or permanent effects. The failure to find a liaison brain in a nonhuman brain might simply indicate that such brains are not conscious. On the other hand, if an animal is conscious, it might be considered just as unethical to experiment on its brain as on one of ours. So we may be eternally barred from the decisive experiment. For all these reasons it will probably always be an option to believe that dualistic gaps exist somewhere, but cannot be observed. However, failing to find the causal chain from brain to mind and back is not what the dualist must worry about. The real threat is that neuroscientists will find another, *physical* chain of events between the tasting of the wine and the uttering of the words.

Suppose we open up the brain of a wine taster and trace out exactly all the neural pathways involved in recognizing and appreciating a sip of wine. Let’s suppose that we have a complete neuroscientific, computational explanation of why she utters “magnifique!” instead of “sacre bleu!” at a key moment. Even if there is an undetected dualist gap, there is nothing for it to do. Nothing that occurs in the gap could affect what she says, because we have (we are imagining) a *complete* explanation of what she says.

One further option for the dualist is to say that experience is a non-physical event-series that accompanies and mirrors the physical one. It plays no causal role, it just happens. This position is called *epiphenomenalism*. This is a possibility, but an unappealing one. For one thing, it seems like an odd way for the universe to be. Why is it that the sort of arrangement of molecules that we find in the brain is accompanied by this extra set of events? Why should the experiences mirror what the

brain does so closely? Keep in mind that the dualist holds that there is no physical explanation of this link, so it is difficult to point to physical properties of the molecules in the brain that would have anything to do with it. The interior of the sun consists of molecules that move around in complex ways. Are these motions accompanied by experiences? When an ice cube melts, its molecules speed up in reaction to the heat of the environment. Does the ice cube feel heat? It can't tell us that it does, but then you can't tell us either. You may *believe* that your utterance of the words "It's hot in here" has something to do with your experience of heat, but it actually depends on various neurophysiological events that are no different in principle from what happens to the ice cube. The experience of heat is something else that happens, on the side. Difficulties such as these make epiphenomenalism almost useless as a framework for understanding consciousness, and I will have little to say about it.

Most scientists and philosophers find the problems with dualism insurmountable. The question is what to replace it with. A solution to the problem of consciousness, or the *mind-body* problem, would be a purely physical mechanism in the brain whose behavior we could identify with having experience. This is a tall order. Suppose we filled it, by finding that mechanism. As we peered at it, we could say with certainty that a certain set of events was the experience of red, that another set was a kind of pleasure, another an excruciating pain. And yet the events would not be fundamentally different from the events we have already observed in brains.

Furthermore, if the brain really is just an organic information-processing system, then the fact that the events occur in neurons would just be a detail. We would expect that if we replaced some or all of the neurons with equivalent artificial systems, the experiences wouldn't change. This seems implausible at first. We think of living tissue as being intrinsically sensitive in ways that silicon and wire could never be. But that intuition is entirely dualistic; we picture living tissue as "exuding" experience in some gaseous manner that we now see isn't right. At a fine enough resolution, the events going on in cells are perfectly mechanical. The wetness disappears, as it were, and we see little molecular machines pulling themselves along molecular ropes, powered by molecular springs. At the level of neurons, these little machines process information, and

they could process the information just as well if they were built using different molecules.

This idea seems so unpalatable to some that they refuse to speculate any further. Consciousness is evidently so mysterious that we will never understand it. We can't imagine any way to build a bridge between physics and mental experience, so we might as well give up. (Colin McGinn is the philosopher most closely associated with this position; see McGinn 1991.) This is an odd position to take when things are just starting to get interesting. In addition, one can argue that it is irresponsible to leave key questions about the human mind dangling when we might clear them up.

Modern culture is in an awkward spot when it comes to the mind-body problem. Scientists and artists are well aware that dualism has failed, but they have no idea what to replace it with. Meanwhile, almost everyone else, including most political and religious leaders, take dualism for granted. The result is that the intellectual elite can take comfort in their superiority over ordinary people and their beliefs, but not in much else. Is this a state of affairs that can last indefinitely without harmful consequences?

I realize that there is a cynical view that people have always accepted delusions and always will. If most people believe in astrology, there can't be any additional harm in their having incoherent beliefs about what goes on inside their heads. I suppose that if you the view the main purpose of the human race as the consumption of products advertised on TV, then their delusions are not relevant. I prefer to think that, at the very least, humans ought to have a chance at the dignity that comes from understanding and accepting their world. Our civilization ought to be able to arrive at a framework in which we appreciate human value without delusions.

### **Nondualist Explanations of Consciousness**

The field of consciousness studies has been quite busy lately. There seem to be two major camps on the mind-body problem: those who believe that we already have the tools we need to explain the mind, and those who believe that we don't and perhaps never will. McGinn is in the pessimistic camp, as is Nagel (1975) and others. I'm an optimist.

Some theorists (notably Bernard Baars 1988, 1996) focus on explaining the function of consciousness and how this function might be realized by the structures of the brain. Although this is an important area, it won't be my main focus. I agree with Chalmers's assessment (Chalmers 1996) that the "hard problem" of consciousness is to explain how a physical object can have experiences. This is the problem of *phenomenal consciousness*. It is a hard problem for all theories, but especially for computational ones.

Like everyone else, I can't define "phenomenal consciousness." It's the ability to have experiences. I assume that anyone who can read this knows what it means to experience something (from their own experience!); and that everyone knows that thermostats don't have experiences, even though they can react to temperature differences.

The difficulty of defining consciousness has led some to propose that there is no thing as consciousness. The standard citations are to Churchland (1990) and Rorty (1965). This position is called *eliminativism*. The idea is to replace the concept of consciousness with more refined (more scientific?) concepts, much as happened with concepts like "energy" and "mass" in past scientific revolutions. It seems plain that a full understanding of the mind will involve shifts of this kind. If we ever do achieve fuller understanding (which the pessimists doubt), any book written before the resulting shift, including this one, will no doubt seem laughably quaint. However, we can't simply wait around for this to happen. We have to work on the problems we see now, using the tools at hand. There is clearly a problem of how a thing, a brain or computer, can have experiences, or appear to have what people are strongly tempted to call experiences. To explain how this is conceivable at all must be our goal.

O'Brien and Opie (1999) make a useful distinction between *vehicle* and *process* theories of phenomenal consciousness:

Either consciousness is to be explained in terms of the nature of the representational vehicles the brain deploys, or it is to be explained in terms of the computational processes defined over these vehicles. We call versions of these two approaches VEHICLE and PROCESS theories of consciousness, respectively.

A process theory is one that explains experience in terms of things the brain does, especially things it computes. A vehicle theory explains experience as a property of the entity doing the computing. O'Brien and Opie themselves propose a vehicle theory in which experience is identified with

(or correlated with?) stable activation patterns in networks of neurons. Another example, much vaguer, is that of John Searle (1992), who appeals to unknown “causal powers” of brain tissue to explain experience.

Then there is the theory of Stuart Hammeroff (1994; Penrose 1994), which explains consciousness in terms of the quantum-mechanical behavior of brain cells, specifically their microtubules. I don’t know how this approach fits into O’Brien and Opie’s dichotomy, because the details are so fuzzy. But if the theory is a process theory, it’s a theory of an unusual kind, because the processes in question cannot, by hypothesis, be explained mechanistically. Instead, the proposal seems to be that the ability of the mind to arrive at sudden insights is explained in the same way as the sudden collapse of wave functions in quantum mechanics. Grant that, and perhaps you may be willing to grant that phenomenal consciousness gets explained somehow, too.

The problem with all these theories besides their vagueness is that they are vulnerable to subversion by competing process theories (McDermott 1999). This is a flaw they share with dualism. As I explained above, any dualistic or “vehicular” explanation of the mind will have to accommodate all the facts that mundane process models explain, a set of facts that one can expect to grow rapidly. Every time we explain a mental ability using ordinary computational processes, we will have to redraw the boundary between what the vehicle theory accounts for and what the process theory accounts for. There are only two ways for this process to be arrested or reversed: either the vehicle theorists must explain more, or the process theorists must fail to explain very much. In the first case, the vehicle theory must compete with process theory on the process theory’s home field, explaining how particular behaviors and competences can result from implementing a computation in one medium rather than another. That’s hard to picture. In the second case, the vehicle theory will get a tie by default; no one will have explained consciousness, so the vehicle theory will have explained it as well as anyone.

If process theory doesn’t fail, however, vehicle theory will be left in an odd position indeed. Suppose we have two entities  $E_1$  and  $E_2$  that behave intelligently, converse on many topics, and can tell you about their experiences. Their similarity is explained by the fact that they implement the same computational processes. However, the one implemented using

vehicle  $V_{genuine}$  is, according to the theory, conscious. The other, implemented using vehicle  $V_{bogus}$ , is only apparently conscious. When it talks of its experiences, it's actually making meaningless sounds, which fool only those unfamiliar with the theory. Unfortunately, no matter how elegant the theory is, it won't supply any actual *evidence* favoring one vehicle over the other. By hypothesis, everything observable about the two systems is explained by the computational process they both instantiate. For instance, if you wonder why  $E_1$  is sometimes unconscious of its surroundings when deeply involved in composing a tune, whatever explanation you arrive at will work just fine for  $E_2$ , because they implement exactly the same computational processes, except that in the case of  $E_2$  you'll have to say, "It's 'apparently conscious' of its surroundings most of the time, except when its working on a new tune, when it's not even apparently conscious."

Vehicle theories are thus likely to be a dead end. This is not to say that the explanation of consciousness may not require new mechanisms. The point is, though, that if they are required they will still be *mechanisms*. That is, they will explain observable events. Phenomenal consciousness is not a secret mystery that is forever behind a veil. When I taste something sour, I purse my lips and complain. A theory must explain why things have tastes, but it must also explain why my lips move in those ways, and the two explanations had better be linked.

Many critics of computational theories reject the idea that phenomenal consciousness can be explained by explaining certain behaviors, such as lip pursing, or utterances such as "Whew, that's sour." But even those who believe that explaining such behavior is insufficient must surely grant that it is *necessary*. We can call this the Principle of the Necessity of Behavioral Explanation: No theory is satisfactory unless it can explain why someone having certain experiences behaves in certain characteristic ways. Naturally, process theories tend to explain behavior well and experience not so well, whereas vehicle theories tend to have the opposite problem.

Process theories tend to fall into two groups, called *first-order* and *higher-order* theories. The former are those in which in some contexts the processing of sensory information is "experience-like" in a way that allows us to say that in those contexts the processing *is* experience. Higher-order theories, on the other hand, say that to be an experience a piece of

sensory information must itself be the object of perception. First-order theories include those of Kirk (1994) and Tye (1995).

Shapiro (2000) sketches what might be considered the “primordial” first-order theory:

Why not think of our perceptual experiences as sometimes entering a channel that makes us phenomenally aware of what they represent and other times bypassing this channel? When I’m driving around town thinking about a lecture I need to prepare, my perceptions of the scenery bypass the phenomenal awareness channel. When, on the other hand, I need to attend more closely to the world, to heighten my awareness of the world, I elect (perhaps unconsciously) to make my perceptions of the world conscious. Accordingly, my perceptions of the world get funneled through the phenomenal awareness channel. Phenomenal awareness, on this picture, requires no higher-order representational capacities.

He says this model is pure speculation, so I don’t suppose he would defend it if pressed, but it makes a convenient target in that its obvious flaw is shared by all first-order theories. The flaw is that nothing is said about what makes events in one channel conscious and those in another channel unconscious. In Kirk’s version, sensory data that are presented to the “main assessment processes” (1994, p. 146) are experienced, whereas other sensory data are not. Why not, exactly? Just as for dualist theories and vehicle theories, first-order computational theories suffer from a disconnection between the hypothesized process and its visible effects. The theories tend to treat “conscious” as a label stuck onto some signals in the brain, without any explanation of how this labeling causes people to be able to report on those signals (while being unable to report on those without the label).

Second-order theories do not have this flaw, because they say exactly what the labels consist in: computational events whose topic is the sensory processes being labeled. Here the labels are as concrete as data structures in a computer, so there is no difficulty following the chain of causality. The difficulty is convincing anyone that the resulting system has real experiences.

I will be pursuing a second-order theory in the rest of the book. Chapter 2 is a survey of the present state of research in artificial intelligence. Chapter 3 is a detailed explanation of my theory of consciousness. Chapter 4 deals with various objections, including those alluded to in the previous paragraph. The most serious objections are based on the

observation that a computational theory of mind can't be correct because concepts such as "computer" and "symbol" are ill defined. Chapter 5 deals with this issue. Chapter 6 deals with various consequences of the theory, including the impact on religion and ethics.

One feature that will strike many readers is how little I appeal to neuroscience, unlike a great many recent theories of consciousness (Flanagan 1992; Crick and Koch 1998; Churchland 1986, 1995). One reason neuroscience is so popular is that it has produced some detailed, interesting proposals for how the brain might work. Many people won't be satisfied with an explanation of thinking and consciousness that doesn't ultimately appeal to the facts of neurophysiology.

But another reason is a prejudice that the basic case of phenomenal consciousness is the quivering bit of protoplasm in contact with the cruel world. The brain feels, in the end, because it is made of living, feeling parts. I have the opposite intuition: that feeling has nothing to do with being alive. The great majority of living things never feel anything. When evolution invented feeling, it stumbled onto a phenomenon that can be elicited from a living system, but not just from living systems. A theory of phenomenal consciousness must reflect this neutrality.

Another way to put it is this: Different organisms sense vastly different things. We rely primarily on visual inputs and so we receive completely different data about the world than a bat would at the same point in space and time. But gathering data and having experiences are two different things, and it may be that adding phenomenal consciousness to data processing is always a matter of adding the same simple twist to the system. If we ever find life on other planets, we will probably find that the data they gather from their environment, and the anatomical structures they use to process them, are specific to that environment and their needs; but if they're conscious it will be because of the same trick that our brains use. *What* they're conscious of will be different, but *the way* they're conscious may be the same. The same conclusions would apply to conscious robots.

This book is mainly about philosophical questions, but I confess that I do not always feel comfortable using the usual philosophical tools to approach them. Usually this is owing to differences in training and disposition. I hope those who would feel more at home in a discussion

conducted in the pure philosophical style will nevertheless bear with me, in spite of my neglect of some of the problems and issues that philosophers focus on.

For example, philosophers spend a lot of time arguing about *functionalism*. This term has several meanings. Some people treat it as synonymous with *computationalism*, the doctrine that the mind can be explained entirely in terms of computation. Since I'm defending a version of computationalism, to that extent I'm defending functionalism, too. However, there are also other meanings assigned to the term, which reduces its utility. One version may be summarized thus: what mental terms and predicates refer to is whatever fills certain causal roles in a functional description of the organism involved. For example, a statement such as "Fred is in pain" is supposed to mean, "Fred is in state X, where X is a state with the properties pain is supposed to have in a worked-out functional description of Fred, e.g., the property of causing Fred to avoid what he believes causes X." Actually, stating the full meaning requires replacing "believe" with a functionally described Y, and so forth.

The purpose of this version of functionalism is to show that, in principle, mental terms can be defined so that they can be applied to systems without making any assumptions about what those systems are made of. If pain can be defined "functionally," then we won't be tempted to define it in terms of particular physical, chemical, or neurological states. So when we find an alien staggering from its crashed spaceship and hypothesize that it is in pain, the claim won't be refutable by observing that it is composed of silicon instead of carbon.

I am obviously in sympathy with the motivation behind this project. I agree with its proponents that the being staggering from the spaceship might be in pain in spite of being totally unlike earthling animals. The question is whether we gain anything by clarifying the definitions of terms. We have plenty of clearcut mental states to study, and can save the borderline cases for later. Suppose one had demanded of Van Loewenhock and his contemporaries that they provide a similar sort of definition for the concept of life and its subconcepts, such as respiration and reproduction. It would have been a complete waste of time, because what Van Loewenhock wanted to know, and what we are now figuring out, is *how life works*. We know there are borderline cases, such as viruses,

but we don't care exactly where the border lies, because our understanding encompasses both sides. The only progress we have made in defining "life" is to realize that it doesn't need to be defined. Similarly, what we want to know about minds is *how they work*. My guess is that we will figure that out, and realize that mental terms are useful and meaningful, but impossible to define precisely.

In practice people adopt what Dennett (1978a) calls the "intentional stance" toward creatures that seem to think and feel. That is, they simply *assume* that cats, dogs, and babies have beliefs, desires, and feelings roughly similar to theirs as long as the assumption accounts for their behavior better than any other hypothesis can. If there ever are intelligent robots, people will no doubt adopt the intentional stance toward them, too, regardless of what philosophers or computer scientists say about the robots' true mental states. Unlike Dennett, I don't think the success of the intentional stance settles the matter. If a system seems to act intentionally, we have to explain *why* it seems that way using evidence besides the fact that a majority of people agree that it does. People are right when they suppose babies have mental states and are wrong when they suppose the stars do.

So I apologize for not spending more time on issues such as the structure of reductionism, the difference between epistemological and metaphysical necessity, and the varieties of supervenience. I am sure that much of what I say could be said (and has been said) more elegantly using those terms, but I lack the requisite skill and patience. My main use of the philosophical literature is the various ingenious thought experiments ("intuition pumps," in Dennett's phrase) that philosophers have used in arguments. These thought experiments tend to have vivid, intuitively compelling consequences; that's their whole purpose. In addition, the apparent consequences are often completely wrong. I believe in those cases it is easy to show that they are wrong without appeal to subtle distinctions; if those familiar with the philosophical intricacies are not satisfied, there are plenty of other sources where they can find the arguments refuted in the correct style. In particular, Daniel Dennett (1991), David Rosenthal (1986, 1993), Thomas Metzinger (1995a), and William Lycan (1987, 1996) defend positions close to mine in philosophers' terms, though they each disagree with me on several points.

Several other nonphilosophers have proposed second-order theories of consciousness. Marvin Minsky is especially explicit about the role of self-models in consciousness (Minsky 1968). Douglas Hofstadter's proposals (Hofstadter 1979; Hofstadter and Dennett 1981) are less detailed, but in many ways more vivid and convincing. Michael Gazzaniga (1998) bases his ideas on neuroscience and psychology.

Adding another voice to this chorus may just confuse matters; but perhaps it may persuade a few more people, and perhaps even clarify the issues a bit.