Chapter 3

Breaking the Hold: Silicon Brains, Conscious Robots, and Other Minds

---

The view of the world as completely objective has a very powerful hold on us, though it is inconsistent with the most obvious facts of our experiences. As the picture is false, we ought to be able to break the hold. I don't know any simple way to do that. One of the many aims of this book, however, is to begin the task. In this chapter I want to describe some thought experiments that will challenge the accuracy of the picture. Initially the aim of the thought experiments is to challenge the conception of the mental as having some important internal connection to behavior.

To begin undermining the foundations of this whole way of thinking, I want to consider some of the relationships between consciousness, behavior, and the brain. Most of the discussion will concern conscious mental phenomena; but leaving out the unconscious at this point is not such a great limitation, because, as I will argue in detail in chapter 7, we have no notion of an unconscious mental state except in terms derived from conscious states. To begin the argument, I will employ a thought experiment that I have used elsewhere (Searle 1982). This *Gedankenexperiment* is something of an old chestnut in philosophy, and I do not know who was the first to use it. I have been using it in lectures for years, and I assume that anybody who thinks about these topics is bound to have something like these ideas occur to him or her eventually.

I. Silicon Brains

Here is how it goes. Imagine that your brain starts to deteriorate in such a way that you are slowly going blind.

Imagine that the desperate doctors, anxious to alleviate your condition, try any method to restore your vision. As a last resort, they try plugging silicon chips into your visual cortex. Imagine that to your amazement and theirs, it turns out that the silicon chips restore your vision to its normal state. Now, imagine further that your brain, depressingly, continues to deteriorate and the doctors continue to implant more silicon chips. You can see where the thought experiment is going already: in the end, we imagine that your brain is entirely replaced by silicon chips; that as you shake your head, you can hear the chips rattling around inside your skull. In such a situation there would be various possibilities. One logical possibility, not to be excluded on any a priori grounds alone, is surely this: you continue to have all of the sorts of thoughts, experiences, memories, etc., that you had previously; the sequence of your mental life remains unaffected. In this case, we are imagining that the silicon chips have the power not only to duplicate your input-output functions, but also to duplicate the mental phenomena, conscious and otherwise, that are normally responsible for your input-output functions.

I hasten to add that I don't for a moment think that such a thing is even remotely empirically possible. I think it is empirically absurd to suppose that we could duplicate the causal powers of neurons entirely in silicon. But that is an empirical claim on my part. It is not something that we could establish a priori. So the thought experiment remains valid as a statement of logical or conceptual possibility.

But now let us imagine some variations on the thought experiment. A second possibility, also not to be excluded on any a priori grounds, is this: as the silicon is progressively implanted into your dwindling brain, you find that the area of your conscious experience is shrinking, but that this shows no effect on your external behavior. You find, to your total amazement, that you are indeed losing control of your external behavior. You find, for example, that when the doctors test your vision, you hear them say, "We are holding up a red object in front of you; please tell us what you see." You want to cry out, "I can't see anything. I'm going totally blind." But

you hear your voice saying in a way that is completely out of your control, "I see a red object in front of me." If we carry this thought experiment out to the limit, we get a much more depressing result than last time. We imagine that your conscious experience slowly shrinks to nothing, while your externally observable behavior remains the same.

It is important in these thought experiments that you should always think of it from the first-person point of view. Ask yourself, "What would it be like for me?" and you will see that it is perfectly conceivable for you to imagine that your external behavior remains the same, but that your internal conscious thought processes gradually shrink to zero. From the outside, it seems to observers that you are just fine, but from the inside you are gradually dying. In this case, we are imagining a situation where you are eventually mentally dead, where you have no conscious mental life whatever, but your externally observable behavior remains the same.

It is also important in this thought experiment to remember our stipulation that you are becoming unconscious but that your behavior remains unaffected. To those who are puzzled how such a thing is possible, let us simply remind them: As far as we know, the basis of consciousness is in certain specific regions of the brain, such as, perhaps, the reticular formation. And we may suppose in this case that these regions are gradually deteriorating to the point where there is no consciousness in the system. But we further suppose that the silicon chips are able to duplicate the input-output functions of the whole central nervous system, even though there is no consciousness left in the remnants of the system.

Now consider a third variation. In this case, we imagine that the progressive implantation of the silicon chips produces no change in your mental life, but you are progressively more and more unable to put your thoughts, feelings, and intentions into action. In this case, we imagine that your thoughts, feelings, experiences, memories, etc., remain intact, but your observable external behavior slowly reduces to total paralysis. Eventually you suffer from total paralysis, even though your mental life is unchanged. So in this case, you might hear the doctors saying,

> The silicon chips are able to maintain heartbeat, respiration, and other vital processes, but the patient is obviously brain dead. We might as well unplug the system, because the patient has no mental life at all.

Now in this case, you would know that they are totally mistaken. That is, you want to shout out,

> No, I'm still conscious! I perceive everything going on around me. It's just that I can't make any physical movement. I've become totally paralyzed.

The point of these three variations on the thought experiment is to illustrate the *causal* relationships between brain processes, mental processes, and externally observable behavior. In the first case, we imagined that the silicon chips had causal powers equivalent to the powers of the brain, and thus we imagined that they caused both the mental states and the behavior that brain processes normally cause. In the normal case, such mental states mediate the relationship between input stimuli and output behavior.

In the second case, we imagined that the mediating relationship between the mind and the behavior patterns was broken. In this case, the silicon chips did not duplicate the causal powers of the brain to produce conscious mental states, they only duplicated certain input-output functions of the brain. The underlying conscious mental life was left out.

In the third case, we imagined a situation where the agent had the same mental life as before, but in this case, the mental phenomena had no behavioral expression. Actually, to imagine this case we need not even have imagined the silicon chips. It would have been very easy to imagine a person with the motor nerves cut in such a way that he or she was totally paralyzed, while consciousness and other mental phenomena remained unaffected. Something like these cases exists in clinical reality. Patients who suffer from the Guillain-Barre syndrome are completely paralyzed, but also fully conscious.

What is the philosophical significance of these three thought experiments? It seems to me there is a number of lessons to

be learned. The most important is that they illustrate something about the relationship between mind and behavior. What exactly is the importance of behavior for the concept of mind? *Ontologically speaking, behavior, functional role, and causal relations are irrelevant to the existence of conscious mental phenomena. Epistemically,* we do learn about other people's conscious mental states in *part* from their behavior. *Causally,* consciousness serves to mediate the causal relations between input stimuli and output behavior; and from an *evolutionary* point of view, the conscious mind functions causally to control behavior. But *ontologically* speaking, the phenomena in question can exist completely and have all of their essential properties independent of any behavioral output.

Most of the philosophers I have been criticizing would accept the following two propositions:

1. Brains cause conscious mental phenomena.
2. There is some sort of conceptual or logical connection between conscious mental phenomena and external behavior.

But what the thought experiments illustrate is that these two cannot be held consistently with a third:

3. The capacity of the brain to cause consciousness is conceptually distinct from its capacity to cause motor behavior. A system could have consciousness without behavior and behavior without consciousness.

But given the truth of 1 and 3, we have to give up 2. So the first point to be derived from our thought experiments is what we might call "the principle of the independence of consciousness and behavior." In case number two, we imagined the circumstance in which the behavior was unaffected, but the mental states disappeared, so behavior is not a sufficient condition for mental phenomena. In case number three, we imagined the circumstance in which mental phenomena were present, but the behavior disappeared, so behavior is not a necessary condition for the presence of the mental either.

Two other points are illustrated by the thought experiments. First, the ontology of the mental is essentially a first-person ontology. That is just a fancy way of saying that every mental state has to be *somebody's* mental state. Mental states only exist as subjective, first-person phenomena. And the other point related to this is that, epistemically speaking, the first-person point of view is quite different from the third-person point of view. It is easy enough to imagine cases, such as those illustrated by our thought experiments, where from a third-person point of view, somebody might not be able to tell whether I had any mental states at all. He might even think I was unconscious, and it might still be the case that I was completely conscious. From the first-person point of view, there is no question that I am conscious, even if it turned out that third-person tests were not available.

II. Conscious Robots

I want to introduce a second thought experiment to buttress the conclusions provided by the first. The aim of this one, as with the first, is to use our intuitions to try to drive a wedge between mental states and behavior. Imagine that we are designing robots to work on a production line. Imagine that our robots are really too crude and tend to make a mess of the more refined elements of their task. But imagine that we know enough about the electrochemical features of human consciousness to know how to produce robots that have a rather low level of consciousness, and so we can design and manufacture conscious robots. Imagine further that these conscious robots are able to make discriminations that unconscious robots could not make, and so they do a better job on the production line. Is there anything incoherent in the above? I have to say that according to my "intuitions," it is perfectly coherent. Of course, it is science fiction, but then, many of the most important thought experiments in philosophy and science are precisely science fiction.

But now imagine an unfortunate further feature of our conscious robots: Suppose that they are absolutely miserable.

Again, we can suppose that our neurophysiology is sufficient for us to establish that they are extremely unhappy. Now imagine we give our robotics research group the following task: Design a robot that will have the capacity to make the same discriminations as the conscious robots, but which will be totally unconscious. We can then allow the unhappy robots to retire to a more hedonically satisfying old age. This seems to me a well-defined research project; and we may suppose that, operationally speaking, our scientists try to design a robot with a "hardware" that they know will not cause or sustain consciousness, but that will have the same input-output functions as the robot that has a "hardware" that does cause and sustain consciousness. We might suppose then that they succeed, that they build a robot that is totally unconscious, but that has behavioral powers and abilities that are absolutely identical with those of the conscious robot.

The point of this experiment, as with the earlier ones, is to show that as far as the ontology of consciousness is concerned, behavior is simply irrelevant. We could have *identical behavior* in two different systems, one of which is conscious and the other totally unconscious.

III. Empiricism and the "Other Minds Problem"

Many empirically minded philosophers will be distressed by these two thought experiments, especially the first. It will seem to them that I am alleging the existence of empirical facts about the mental states of a system that are not ascertainable by any empirical means. Their conception of the empirical means for ascertaining the existence of mental facts rests entirely on the presupposition of behavioral evidence. They believe that the only evidence we have for attributing mental states to other systems is the behavior of those systems.

In this section I want to continue the discussion of the other minds problem that was begun in chapter 1. Part of my aim will be to show that there is nothing incoherent or objectionable in the epistemic implications of the two thought experiments I just described, but my primary aim will be to give an

account of the "empirical" basis we have for supposing that other people and higher animals have conscious mental phenomena more or less like our own.

It is worth emphasizing at the beginning of the discussion that in the history of empirical philosophy and of the philosophy of mind, there is a systematic ambiguity in the use of the word "empirical," an ambiguity between an ontological sense and an epistemic sense. When people speak of empirical facts, they sometimes mean actual, contingent facts in the world as opposed to, say, facts of mathematics or facts of logic. But sometimes when people speak of empirical facts, they mean facts that are testable by third-person means, that is, by "empirical facts" and "empirical methods," they mean facts and methods that are accessible to all competent observers. Now this systematic ambiguity in the use of the word "empirical" suggests something that is certainly false: that all empirical facts, in the ontological sense of being facts in the world, are equally accessible epistemically to all competent observers. We know independently that this is false. There are lots of empirical facts that are not equally accessible to all observers. The previous sections gave us some thought experiments designed to show this, but we actually have empirical data that suggest exactly the same result.

Consider the following example) We can with some difficulty imagine what it would be like to be a bird flying. I say "with some difficulty" because, of course, the temptation is always to imagine what it would be like *for us* if we were flying, and not, strictly speaking, what it is like for *a bird* to be flying. But now some recent research tells us that there are some birds that navigate by detecting the earth's magnetic field. Let us suppose that just as the bird has a conscious experience of flapping its wings or feeling the wind pressing against its head and body, so it also has a conscious experience of a feeling of magnetism surging through its body. Now, what is it like to feel a surge of magnetism? In this case, I do not have the faintest idea what it feels like for a bird, or for that matter, for a human to feel a surge of magnetism from the

earth's magnetic field. It is, I take it, an empirical fact whether or not birds that navigate by detecting the magnetic field actually have a conscious experience of the detection of the magnetic field. But the exact qualitative character of this empirical fact is not accessible to standard forms of empirical tests. And indeed, why should it be? Why should we assume that all the facts in the world are equally accessible to standard, objective, third-person tests? If you think about it, the assumption is obviously false.

I said that this result is not as depressing as it might seem. And the reason is simple. Although in some cases we do not have equal access to certain empirical facts because of their intrinsic subjectivity, in general we have indirect methods of getting at the same empirical facts. Consider the following example. I am completely convinced that my dog, as well as other higher animals, has conscious mental states, such as visual experiences, feelings of pain, and sensations of thirst and hunger, and of cold and heat. Now why am I so convinced of that? The standard answer is because of the dog's behavior, because by observing his behavior I infer that he has mental states like my own. I think this answer is mistaken. It isn't just because the dog behaves in a way that is appropriate to having conscious mental states, but also because I can see that the causal basis of the behavior in the dog's physiology is relevantly like my own. It isn't just that the dog has a structure like my own and that he has behavior that is interpretable in ways analogous to the way that I interpret my own. But rather, it is in the combination of these two facts that I can see that the behavior is appropriate and that it has the appropriate *causation* in the underlying physiology. I can see, for example, that these are the dog's ears; this is his skin; these are his eyes; that if you pinch his skin, you get behavior appropriate to pinching skin; if you shout in his ear, you get behavior appropriate to shouting in ears.

It is important to emphasize that I don't need to have a fancy or sophisticated anatomical and physiological theory of dog structure, but simple, so to speak, "folk" anatomy and

physiology—the ability to recognize the structure of skin, eyes, teeth, hair, nose, etc., and the ability to suppose that the causal role that these play in his experiences is relevantly like the causal role that such features play in one's own experiences. Indeed, even describing certain structures as "eyes" or "ears" already implies that we are attributing to them functions and causal powers similar to our own eyes and ears. In short, though I don't have direct access to the dog's consciousness, nonetheless it seems to me a well-attested empirical fact that dogs are conscious, and it is attested by evidence that is quite compelling. I do not have anything like this degree of confidence when it comes to animals much lower on the phylogenetic scale. I have no idea whether fleas, grasshoppers, crabs, or snails are conscious. It seems to me that I can reasonably leave such questions to neurophysiologists. But what sort of evidence would the neurophysiologist look for? Here, it seems to me, is another thought experiment that we might well imagine.

Suppose that we had an account of the neurophysiological basis of consciousness in human beings. Suppose that we had quite precise, neurophysiologically isolable causes of consciousness in human beings, such that the presence of the relevant neurophysiological phenomena was both necessary and sufficient for consciousness. If you had it, you were conscious; if you lost it, you became unconscious. Now imagine that some animals have this phenomenon, call it "x" for short, and others lack it. Suppose that x was found to occur in all those animals, such as ourselves, monkeys, dogs, etc., of which we feel quite confident that they are conscious on the basis of their gross physiology, and that x was totally absent from animals, such as amoebae, to which we do not feel inclined to ascribe any consciousness. Suppose further that the removal of x from any human being's neurophysiology immediately produced unconsciousness, and its reintroduction produced consciousness. In such a case, it seems to me we might reasonably assume that the presence of x played a crucial causal role in the production of consciousness, and this discovery would enable us to settle doubtful cases of animals either having or

lacking conscious states. If snakes had *x,* and mites lacked it, then we might reasonably infer that mites were operating on simple tropisms and snakes had consciousness in the same sense that we, dogs, and baboons do.

I don't for a moment suppose that the neurophysiology of consciousness will be as simple as this. It seems to me much more likely that we will find a great variety of forms of neurophysiologies of consciousness, and that in any real experimental situation we would seek independent evidence for the existence of mechanical-like tropisms to account for apparently goal-directed behavior in organisms that lacked consciousness. The point of the example is simply to show that we can have indirect means of an objective, third-person, empirical kind for getting at empirical phenomena that are intrinsically subjective and therefore inaccessible to direct third-person tests.

It shouldn't be thought, however, that there is something second rate or imperfect about the third-person empirical methods for discovering these first-person subjective empirical facts. The methods rest on a rough-and-ready principle that we use elsewhere in science and in daily life: *same causes-same effects,* and *similar causes-similar effects.* We can readily see in the case of other human beings that the causal bases of their experiences are virtually identical with the causal bases of our experiences. This is why in real life there is no "problem of other minds." Animals provide a good test case for this principle because, of course, they are not physiologically identical with us, but they are in certain important respects similar. They have eyes, ears, nose, mouth, etc. For this reason we do not really doubt that they have the experiences that go with these various sorts of apparatus. So far, all these considerations are prescientific. But let us suppose that we could identify for the human cases exact causes of consciousness, and then could discover precisely the same causes in other animals. If so, it seems to me we would have established quite conclusively that other species have exactly the same sort of consciousness that we have, because we can presume that the same causes produce the same effects. This would not be just a wild speculation, because we would have very good reason to

suppose that those causes would produce the same effects in other species.

In actual practice, neurophysiology textbooks routinely report, for example, how the cat's perception of color is similar to and different from the human's *and even other animals.* What breathtaking irresponsibility! How could the authors pretend to have solved the other cat's mind problem so easily? The answer is that the problem is solved for cats' vision once we know exactly how the cat's visual apparatus is similar to and different from our own and other species'.2

Once we understand the causal basis of the ascription of mental states to other animals, then several traditional skeptical problems about "other minds" have an easy solution. Consider the famous problem of spectrum inversion that I mentioned in chapter 2. It is often said that, for all we know, one section of the population might have a red /green inversion such that though they make the same behavioral discriminations as the rest of us, the actual experiences they have when they see green, and which they call "seeing green," are experiences that we would, if we had them, call "seeing red," and vice versa. But now consider: Suppose we actually found that a section of the population actually did have the red and green receptors reversed in such a way, and so connected with the rest of their visual apparatus, that we had overwhelming neurophysiological evidence that though their molar discriminations were the same as ours, they actually had different experiences underlying them. This would not be a problem in philosophical skepticism, but a well-defined neurophysiological hypothesis. But then if there is no such section of the population, if all of the non-color-blind people have the same red/green perceptual pathways, we have solid empirical evidence that things look to other people the way they look to us. A cloud of philosophical skepticism condenses into a drop of neuroscience.

Notice that this solution to "the other minds problem," one that we use in science and in daily life, gives us sufficient but not necessary conditions for the correct ascription of mental phenomena to other beings. We would, as I suggested earlier

in this chapter, need a much richer neurobiological theory of consciousness than anything we can now imagine to suppose that we could isolate necessary conditions of consciousness. I am quite confident that the table in front of me, the computer I use daily, the fountain pen I write with, and the tape-recorder I dictate into are quite unconscious, but, of course, I cannot *prove* that they are unconscious and neither can anyone else.

IV. Summary

In this chapter I have so far had two objectives: First, I have tried to argue that as far as the ontology of the mind is concerned, behavior is simply irrelevant. Of course in real life our behavior is crucial to our very existence, but when we are examining the existence of our mental states as mental states, the correlated behavior is neither necessary nor sufficient for their existence. Second, I have tried to begin to break the hold of three hundred years of epistemological discussions of "the other minds problem," according to which behavior is the sole basis on which we know of the existence of other minds. This seems to me obviously false. It is only because of the *connection* between behavior and the causal structure of other organisms that behavior is at all relevant to the discovery of mental states in others.

A final point is equally important: except when doing philosophy, there really is no "problem" about other minds, because we do not hold a "hypothesis," "belief," or "supposition" that other people are conscious, and that chairs, tables, computers, and cars are not conscious. Rather, we have certain Background ways of behaving, certain Background capacities, and these are constitutive of our relations to the consciousness of other people. It is typical of philosophy that skeptical problems often arise when elements of the Background are treated as if they were hypotheses that have to be justified. I don't hold a "hypothesis" that my dog or my department chairman is conscious, and consequently the question doesn't arise except in philosophical debate.