

THE MACHINE QUESTION

On the Possibility of Consciousness in Inorganic Systems

Jeffrey P. Johnson

Independent Study in Philosophy

University of Lancaster, 1986-87

Supervisor: Vernon Pratt

I lost the original of this essay. I recreated it in 2026 from my memory, the reading list, and my surviving college papers, using only sources that were in print by early 1987. The citations have been checked against the editions named in the notes.

"...if there were machines which bore a resemblance to our body and imitated our actions as far as it was morally possible to do so, we should always have two very certain tests by which to recognise that, for all that, they were not real men."

Descartes, *Discourse on the Method*, Part V (1637)

"I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted."

Turing, "Computing Machinery and Intelligence" (1950)

1. Introduction: The Question and the Wager

Is there anything about consciousness that requires it to be made of living matter? I will argue that there is not; that nothing we know to date rules out consciousness in an inorganic system, and that the best theory of mind now available actually predicts that such a thing is possible. I do not claim that any machine now in existence is conscious. I claim only that the door is open, and that the arguments offered for shutting it do not succeed. In this section I will state the question more carefully, say why it is worth asking now, set out the assumption I am starting from, and offer a working definition of the thing in question. The rest of the paper follows the plan given at the end of this section.

First of all, the question is not "Can a computer think?" That question has been asked so often, and answered so confidently on both sides, that it has become a kind of slogan. The question I want to ask is wider and, I think, more basic. A computer is only one kind of inorganic system, and it happens to be the kind we know how to build. The real issue is whether the stuff a thing is made of has anything to do with whether that thing can have a mind. If it does, then the computer question is settled before it is asked. If it does not, then the computer question becomes a question about organization rather than about materials, and it is a question we can actually work on.

An example may make the distinction clear. Nobody asks whether a wooden clock can "really" keep time, or whether only a brass one can. A clock made of wood and a clock made of brass both keep time if their parts are arranged in the right way, and keeping time is not a property of brass. It is a property of the arrangement. The question of this paper is whether consciousness is like keeping time in this respect, or whether it is more like rusting, which brass does and wood does not, because rusting is a matter of what a thing is made of and not of how it is put together. I will try to show that everything we can say about consciousness points to the first answer and nothing points to the second.

The question is worth asking now because the machines have started to arrive. In Britain the Alvey Programme has been funding research in artificial intelligence for four years, in answer to the Japanese Fifth Generation project; expert systems are in commercial use; and this year two volumes on "parallel distributed processing" have appeared which propose that a machine should be organized the way a brain is, out of many simple units working at once, rather than as a single processor following a list of rules.¹ I have spent this year writing programs in Lisp and Prolog for the artificial intelligence course here, and I can report that none of them is conscious. But I can also report that the reasons they are not conscious have to do with what they do, and not with the silicon they run on. That observation is the seed of this paper.

I should say at the outset what I am assuming, since a reader who does not share the assumption will want to stop here. I am assuming that a soul is not required for consciousness. I do not argue for this assumption; I take it as the starting point, in the way that Sartre takes it when he says that existence precedes essence.² Sartre's example is a paper-knife. The artisan who makes a paper-knife has the idea of it first, its purpose and the way it will be made, and the knife exists afterwards, in order to fit the idea. Its essence precedes its existence. Sartre's claim about human beings is that there is no such artisan and no such prior idea: we exist first, and what we are is nothing more than what we make of ourselves. I accept this, and I want to notice something about it that Sartre does not pursue. If there is no prior essence that settles what a human being can become, then there is no prior essence that settles what anything can become. A machine has a designer, it is true, and in that sense it is like the paper-knife. But the designer's idea does not exhaust what the machine is, any more than a parent's idea exhausts what a child is. Whether a thing is conscious is a question about the thing, not about what somebody meant it to be.

There is a second reason for making this assumption openly. If consciousness requires a soul, then the question of this paper is not a question at all; it is answered in advance, and answered by a doctrine that cannot be examined. I would rather ask a question that might turn out to have an answer. The reader who thinks a soul is required may treat everything that follows as an account of what the world would have to be like if he were wrong.

Now I must say what I mean by "consciousness," since the word is used to mean everything from being awake to being wise. I do not think a definition is possible at this stage, and I do not think one is needed. What is needed is a working description that picks out the thing we are asking about, and I will build one out of three ideas that are already on the table.

The first comes from Thomas Nagel. Nagel says that an organism has conscious mental states "if and only if there is something that it is like to be that organism, something it is like for the organism."³ By this he means that consciousness is a matter of there being a point of view. There is something it is like to be me, sitting at this desk; there is, Nagel thinks, something it is like to be a bat, though we cannot say what; and there is nothing it is like to be a stone. This is the part of consciousness that is hardest to talk about and, as I will argue in Section 7, the part that no theory has yet explained. But it is the part that matters, and any working description has to start with it.

The second idea comes from Douglas Hofstadter. In *Gödel, Escher, Bach* Hofstadter argues that a self arises when a system becomes able to represent itself, when among the symbols it manipulates there is a symbol for the system doing the manipulating.⁴ He calls the resulting structure a "strange loop." By this he means that the system's model of the world contains a model of the modeller, so that the system's account of what is going on includes itself as one of the things going on. I think this is the nearest thing we have to an account of

how a point of view could come to exist. A point of view is a place from which things are seen, and a system that models itself has such a place, because it can tell the difference between what is happening to it and what is happening elsewhere.

The third idea comes from Daniel Dennett. Dennett calls something an "intentional system" when we can predict what it will do by supposing that it has beliefs and desires and acts on them.⁵ By this he means that the question whether something has a mind is, in practice, the question whether treating it as having one works. We treat the chess-playing program as wanting to protect its queen because that is the only way to predict its moves; we do not treat the thermostat that way, because "it wants the room warmer" tells us nothing that "the contact closes at sixty-five degrees" does not. Dennett's idea does not tell us whether there is something it is like to be the chess program. But it gives us a way of saying, from the outside, which systems are candidates.

Putting these together, the working description is this: a conscious system is one that has a point of view (Nagel), sustained by a model of itself (Hofstadter), such that the system's behavior is best predicted by ascribing beliefs and desires to it (Dennett). I do not offer this as a definition. I offer it as a description that fits the one clear case, ourselves, and that says nothing whatever about what the system must be made of. That is the point of it. If the description is right, then the question of materials is a separate question, and it has to be argued separately. I have not found anyone who argues it, as opposed to assuming it, and much of this paper is an attempt to find out why.

One more word needs attention, and that is "machine." When Descartes uses the word he means a mechanism, a clockwork, a thing whose parts push on each other; when Turing uses it he means a digital computer; and when people use it in ordinary speech they seem to mean anything made in a factory. I will use it in Turing's sense where I am discussing Turing and Searle, since that is the sense in which they use it, and I will use it in the widest sense, to mean any system that is not alive, when the argument turns in Section 9 to the question whether consciousness might be found outside computers altogether. Where the difference matters I will say which sense I mean.

The plan of the paper is as follows. Section 2 goes back to Descartes, who built the machine and then barred it from having a mind, and shows how Ryle and Wittgenstein removed the bar. Section 3 lays out the positions available to a materialist and settles on functionalism as the working hypothesis. Sections 4 and 5 take up the two tests Descartes proposed for telling a machine from a man: the test of language, which Turing turned into the imitation game and Searle attacked with the Chinese room, and which I will argue is evidence rather than proof either way; and the test of generality, which I have left for Section 8 because I think it is the sharper of the two and the one the machines we have so far actually fail. Section 6 returns to Hofstadter and the origin of the self. Section 7 takes up the objection that no functional story can explain what it is like to be anything, and

concedes as much of it as has to be conceded. Section 9 asks whether the argument, if it works for computers, works for matter in general. Section 10 sums up and makes a prediction, in the spirit of Turing's, which the reader is welcome to hold against me.

I want to end this introduction with an observation from an earlier paper of mine, because it states the attitude I am bringing to the whole question. The laws of nature describe what happens; they do not prescribe what ought to happen. If you throw a rock into the air it comes down, and if one day it did not come down we would not call the rock perverse, we would amend the law.⁶ I think the same holds for the "law" that only living things can be conscious. It is a description of every case we have met so far, and it may be a true description of every case there will ever be. But it is not a prescription, and if a machine were one day to give us good reason to think there was something it was like to be it, the right response would not be to say that this is impossible. It would be to amend the law.

Notes to Section 1

1. David E. Rumelhart, James L. McClelland, and the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 2 vols. (Cambridge, Mass.: MIT Press, 1986). On the Alvey Programme and the Fifth Generation project see Edward Feigenbaum and Pamela McCorduck, *The Fifth Generation* (Reading, Mass.: Addison-Wesley, 1983).
2. Jean-Paul Sartre, *Existentialism and Humanism*, trans. Philip Mairet (London: Methuen, 1948), pp. 26-28.
3. Thomas Nagel, "What Is It Like to Be a Bat?", *Philosophical Review* 83 (1974), p. 436; reprinted in Douglas R. Hofstadter and Daniel C. Dennett, eds., *The Mind's I* (Harmondsworth: Penguin, 1982), p. 392.
4. Douglas R. Hofstadter, *Gödel, Escher, Bach: An Eternal Golden Braid* (Harmondsworth: Penguin, 1980), chapter XX, "Strange Loops, Or Tangled Hierarchies," pp. 684-719, especially "Strange Loops as the Crux of Consciousness," pp. 708-709.
5. Daniel C. Dennett, "Intentional Systems," in *Brainstorms: Philosophical Essays on Mind and Psychology* (Hassocks: Harvester Press, 1979), pp. 3-4.
6. Jeffrey P. Johnson, "Homosexuality: Is It Morally Permissible?", final paper for Philosophy 1003, University of Minnesota, 1985. The example was used there against the argument that a practice can be condemned as "contrary to nature."

Epigraphs: Descartes in the Haldane and Ross translation, *The Philosophical Works of Descartes*, vol. 1 (Cambridge: Cambridge University Press, 1911), p. 116; Turing, *Mind* 59 (1950), p. 442.

2. The Cartesian Inheritance

The question of this paper was invented by Descartes, and it is worth going back to see how he invented it, because he did two things at once that have been pulling against each other ever since. He described the human body as a machine, more thoroughly than anyone had before him, and then he declared that a machine could never have a mind. Both halves of that are still with us. Everyone who says a computer could never think is repeating the second half, and everyone who says the brain is a computer is repeating the first. In this section I will set out what Descartes claimed, then show how Ryle and Wittgenstein took the second half apart, and end by saying what the question looks like once they have finished with it.

Descartes builds the machine

In the fifth part of the *Discourse on the Method* Descartes describes the body, its heart and blood and nerves, as a mechanism, and he does not hedge about it. The body, he says, is a machine "which, having been made by the hands of God, is incomparably better arranged, and possesses in itself movements which are much more admirable, than any of those which can be invented by man."¹ By this he means that the difference between the body and a clock is a difference of workmanship and not of kind. A clock made by God would be a body. He goes on to say that if there were machines with the organs and shape of a monkey, we would have no way of telling them from real monkeys, and this is not a joke at the monkey's expense; it is a statement of what Descartes thinks an animal is. An animal is a machine. The dog that comes running at the sound of its name is doing what a clock does when it strikes the hour, and Descartes says as much of the magpie and the parrot, which can be taught to utter words but cannot, he thinks, be taught to speak.

Then comes the turn. If there were machines that resembled our own bodies and imitated our actions as far as possible, Descartes says, we would still have "two very certain tests" for telling that they were not real men.² The first is that the machine could never use words to reply to whatever is said in its presence, as the dullest man can do. The second is that although it might do some things as well as we do or better, it would fail at others, and so show that it acted "not from knowledge, but only from the disposition of its organs."³ I will come back to both tests, the first in Section 4 and the second in Section 8, because I think they are the two best things ever said against the possibility I am defending, and they deserve a section each. For now the point is simply that Descartes thought he could tell.

Why was he so sure? The answer is in the *Meditations*, and it comes in two steps. The first step is the famous one. In the second Meditation Descartes doubts everything he can, and finds that the one thing he cannot doubt is that he is doubting: "I am, I exist, is

necessarily true each time that I pronounce it, or that I mentally conceive it."⁴ The second step is the one that matters here. Having found that he cannot doubt his own existence as a thinking thing, while he can doubt the existence of his body, he concludes in the sixth Meditation that the two must be different things: this I, "that is to say, my soul by which I am what I am, is entirely and absolutely distinct from my body, and can exist without it."⁵

It would seem to me that this second step is where the trouble starts, and it is worth being clear about what kind of step it is. Descartes has found that he can imagine himself without a body. From this he concludes that he is a thing that could exist without a body. But the fact that I can imagine one thing without another does not show that they are two things. I can imagine the morning star without the evening star, and they are the same planet. What Descartes has actually shown is that his knowledge of his mind and his knowledge of his body come by different routes. He has not shown that there are two things at the ends of the routes. Yet everything that follows in the Cartesian picture depends on there being two: a body, which is a machine, and a mind, which is not, joined somewhere (Descartes suggested the pineal gland) in a way that nobody, including Descartes, was ever able to explain.

I want to notice one more thing before leaving Descartes, because it shows that he felt the difficulty himself. He says that the soul is not lodged in the body "as a pilot in a vessel," but is so closely joined to it that the two form a single whole; otherwise, he says, I would not feel pain when my body is hurt but would merely perceive the damage, as a sailor perceives by sight that his ship is broken.⁶ This is exactly right as a description of what it is like to have a body, and it is exactly what his theory cannot allow. A pilot and a ship are two things, and Descartes has just said that the soul and body are two things. If they are two things then the pilot picture is the right one, and the feeling of pain is left unexplained. If the pilot picture is wrong, as Descartes says it is, then it is hard to see what the "absolute distinction" comes to. He wanted both, and one cannot have both.

Ryle removes the ghost

Gilbert Ryle's *The Concept of Mind* opens with a chapter called "Descartes' Myth," and its target is the picture I have just described, which Ryle calls, "with deliberate abusiveness," the dogma of the Ghost in the Machine.⁷ Ryle's claim is that the picture is not false in the ordinary way, the way it is false that Oxford is north of Lancaster, but false in a special way: it is what he calls a category mistake.

Ryle's example is the visitor who is shown around Oxford. He sees the colleges, the libraries, the playing fields, the offices, and at the end he asks, "But where is the University?"⁸ The visitor's mistake is not that he has missed a building. It is that he thinks the University is another thing of the same kind as the colleges and libraries, when in fact it is the way all of those are organized. Ryle's second example is the foreigner watching

cricket, who is shown the bowlers, the batsmen, the fielders and the umpires, and then asks which of them contributes the team spirit.⁹ Team spirit is not another player. It is a way the players play.

Descartes' mistake, on Ryle's account, is the visitor's mistake. Having described the body and all it does, Descartes asks where the mind is, and since it is not one of the parts, he concludes that it must be another thing of a ghostly kind, invisibly running the machine. Ryle's reply is that the mind is not another thing at all. When we say a man is intelligent, we are not reporting the presence of a second, hidden occupant; we are saying something about how he does what he does, that he does it carefully, with attention to the result, correcting himself as he goes. Intelligence is in the doing, in the way the University is in the colleges.

An example of my own may help. If a man is hurrying along the street, it would be absurd to ask where the hurry is, or what part of him is doing the hurrying while the rest of him walks. The hurry is not a part. It is the manner of the walking. Ryle's claim is that mind stands to behavior as hurry stands to walking, and that Descartes, having looked for the hurry among the parts and not found it, invented an invisible hurrier.

I think Ryle's argument is right as far as it goes, and it goes further than is sometimes allowed. What it does not do, and what Ryle did not claim it did, is explain the thing Descartes was surest of, which is that there is something it is like to be in pain, over and above the wincing. Ryle has shown that the mind is not a second thing. He has not shown that there is nothing left to explain once the behavior is explained. I will come back to this in Section 7. For now the result is a negative one, but it is the negative result this paper needs: whatever a mind is, it is not an extra item that a body either contains or lacks. It is something a body does, or something that goes on in the doing.

Wittgenstein removes the private object

Ryle's argument leaves one place for the ghost to hide, and Wittgenstein went in after it. The place is the inner life. Even if the mind is not a separate thing, it may still seem that each of us has a private world of sensations and thoughts that only he can see into, and that this private world is what consciousness really is. If that is so, then the question about machines becomes the question whether a machine has such a world inside it, and since we cannot see into anyone's world but our own, the question looks unanswerable.

The argument against this picture runs through the sections of the *Philosophical Investigations* on private language.¹⁰ I will not try to reproduce all of it, but the central move can be stated briefly. Wittgenstein asks how a word like "pain" gets its meaning. If its meaning were a private object, something I alone can inspect, then nobody could ever learn the word from anybody else, and nobody could ever check that he was using it the same way twice. But we do learn the word, from our parents, when we cry and are comforted, and we do use it correctly. So its meaning cannot be a private object. Wittgenstein's picture is of a

box that everyone has, into which nobody else can look; everyone says there is a "beetle" in his box.¹¹ By this he means that if the word "beetle" has a use in the language, that use cannot depend on what is in the box, since for all anyone knows the boxes are all different, or empty. The thing in the box "cancels out, whatever it is." The word gets its meaning from the shared use, not from the private object.

This does not mean there are no sensations. Wittgenstein is at pains to say that it does not. What it means is that when we ask whether something has sensations, we are not asking a question about the contents of a box we cannot open. We are asking a question that has public criteria, and Wittgenstein says what they are: "only of a living human being and what resembles (behaves like) a living human being can one say: it has sensations; it sees; is blind; hears; is deaf; is conscious or unconscious."¹² I want to draw attention to the words in parentheses, because they are the hinge of this whole paper. What resembles a living human being, Wittgenstein says, and then he explains what he means by resembling: "behaves like." The criterion is behavior. It is not ingredients. When the vet says the dog is in pain, he has not examined its nerves; he has watched it.

Now consider what Wittgenstein says about machines directly. "Could a machine think? Could it be in pain? Well, is the human body to be called such a machine? It surely comes as close as possible to being such a machine."¹³ And in the next section: "But a machine surely cannot think! Is that an empirical statement? No. We only say of a human being and what is like one that it thinks. We also say it of dolls and no doubt of spirits too. Look at the word 'to think' as a tool."¹⁴ By this he means that "a machine cannot think" is not a discovery about machines. It is a remark about how we use the word "think," and it records the fact that up to now we have only had occasion to use it of people and of things that behave like people. That is a fact about our history, not about what is possible. If a machine came to behave like a person, the remark would simply stop being true, in the way "no one has ever run a mile in under four minutes" stopped being true, without any change in what running is.

I do not want to claim more for Wittgenstein than he claimed. He did not say machines could think. He said the question was not the kind of question people took it to be. And there is a famous sentence in the second part of the *Investigations* that cuts the other way: "If a lion could talk, we could not understand him."¹⁵ By this I take him to mean that understanding is not a matter of decoding words but of sharing a way of life, and that a creature whose life was different enough from ours would remain closed to us even if it used our words. That is a warning worth keeping. A machine that passed every test might still be a lion.

Where this leaves the question

Let me sum up what has happened to Descartes' picture. He gave us a body that is a machine and a mind that is a second thing, joined to the machine in a way he could not explain. Ryle showed that the second thing is a category mistake, an invisible hurrier invented because the hurry could not be found among the parts. Wittgenstein showed that the inner world, the last place the ghost could hide, is not what our words for consciousness refer to, and that the criteria for consciousness are public ones, resemblance and behavior. Between them they have not shown that a machine can be conscious. They have shown something more useful for my purposes: that the question is not what a thing is made of but what it does and how it is organized to do it.

That is the result I need to go forward, and I want to state it as a verdict rather than a hope. Dualism is closed, twice over: by the assumption I made in Section 1, and by the arguments of this section, which do not depend on that assumption. The question of this paper is now a question inside materialism. It asks what organization of matter is enough for consciousness, and whether that organization has to be organic. Section 3 sets out the materialist positions that are available and chooses one. The reader should notice, though, that the ghost has left one thing behind on its way out. Descartes was sure that there is something it is like to be in pain, and neither Ryle nor Wittgenstein denied it. Whatever theory I adopt will have to say something about that, and the honest thing to say now is that none of them says enough.

Notes to Section 2

1. René Descartes, *Discourse on the Method*, Part V, in *The Philosophical Works of Descartes*, trans. Elizabeth S. Haldane and G. R. T. Ross, vol. 1 (Cambridge: Cambridge University Press, 1911), p. 116.
2. Ibid., p. 116.
3. Ibid., p. 116. The Haldane and Ross wording is "not from knowledge, but only from the disposition of their organs."
4. Descartes, *Meditations on First Philosophy*, Meditation II, Haldane and Ross, vol. 1, p. 150.
5. Meditation VI, Haldane and Ross, vol. 1, p. 190.
6. Meditation VI, Haldane and Ross, vol. 1, p. 192.
7. Gilbert Ryle, *The Concept of Mind* (Harmondsworth: Penguin, 1963), p. 17.
8. Ibid., pp. 17-18.
9. Ibid., p. 18.
10. Ludwig Wittgenstein, *Philosophical Investigations*, trans. G. E. M. Anscombe, 3rd ed. (Oxford: Basil Blackwell, 1967), §§243-315.
11. Ibid., §293.
12. Ibid., §281.
13. Ibid., §359.
14. Ibid., §360.

15. Ibid., Part II, p. 223.

3. The Map of Positions

Section 2 closed the door on dualism and left the question inside materialism. But materialism is not one position; it is a family of them, and they disagree about exactly the thing this paper is asking. In this section I will lay out the positions available, using the map Paul Churchland gives in *Matter and Consciousness*, and I will choose one of them as the working hypothesis for the rest of the paper.¹ I will take them in the order Churchland does, since it is also the order in which each was proposed to fix the problems of the one before: behaviorism, then the identity theory, then functionalism, with eliminative materialism as a possibility standing off to one side. My aim is not to prove that functionalism is true. It is to show that it is the position that fits what Section 2 established, and that its main rival fails for a reason that bears directly on machines.

Behaviorism

The first materialist position is the one Ryle is usually taken to hold, and Churchland calls it philosophical behaviorism. On this view, to have a mind is to be disposed to behave in certain ways. To be in pain is to be disposed to wince, cry out, favor the injured part, and say "it hurts" when asked. There is nothing behind the dispositions; the dispositions are the pain.

I have a good deal of sympathy for this view, and Section 2 leaned on its best insight, which is that behavior is the criterion by which we tell that something has a mind. But there is a difference between a criterion and a definition, and behaviorism collapses the two. Putnam's example makes the difference clear. Imagine a community of "super-Spartans," people who have been trained from birth to show no sign of pain whatever, who feel it as we do but never wince, never cry out, never favor the injured part.² They have the pain and lack the behavior. Now imagine the opposite, a good actor, who has the behavior and lacks the pain. If behaviorism were right, the super-Spartan would not be in pain and the actor would be, and both of those are false. What has gone wrong is easy to see. Behavior is how we find out about pain, in the way that smoke is how we find out about fire. But the fire is not the smoke, and a fire that gives off no smoke is still a fire.

Notice that this does not undo Wittgenstein's point in Section 2. Wittgenstein said that "behaves like" is the criterion for saying something has sensations, and that is still right. What Putnam's example shows is that the criterion can be present without the thing and the thing without the criterion, so that the thing must be something more than the criterion. The question is what more.

The identity theory

The second position answers: the thing is a process in the brain. This is the identity theory, proposed by Place and Smart and worked out most fully by Armstrong.³ On this view pain is not a disposition to behave; it is a physical state that causes the behavior, and that state is a state of the brain. Smart's comparison is with lightning. Lightning is an electrical discharge; not "correlated with" one, not "caused by" one, but identical to one, and this was discovered by science, not by analysis of the word "lightning." In the same way, Smart says, pain will turn out to be a brain process, and the discovery will be a scientific one.⁴

This is a materialism that takes the inner state seriously, and it answers the super-Spartan straight off: he is in pain because the brain process is going on, whatever his face is doing. So far it is an improvement. But the identity theory says something more specific than "pain is a physical state." It says pain is a particular kind of physical state, the kind that happens in brains like ours, and it is here that the theory runs into a wall, and the wall is the subject of this paper.

The difficulty was pointed out by Putnam and has come to be called the argument from multiple realizability.⁵ If pain is identical to some particular brain process, call it the firing of C-fibres, then nothing without C-fibres can be in pain. But an octopus has a nervous system built on a quite different plan from ours, and it would be a bold man who said an octopus cannot feel pain. If we ever met an intelligent creature from another planet, with a chemistry unlike ours, we would want to ask whether it felt pain, and the identity theory says the question is already answered in the negative, before we have looked. That cannot be right. The theory has taken a fact about how pain happens to be realized in us and turned it into a definition of pain.

An analogy will show what kind of mistake this is. Suppose someone said that a chair is a certain arrangement of wood. Every chair he had ever sat in was wooden, and the claim would be true of all of them. But then a plastic chair is brought in, and by his definition it is not a chair, though everyone sits in it. What he has done is confuse what chairs are made of with what makes a thing a chair, which is a matter of what it is for and how it does it. Being a chair is a job, and many materials can do the job. The identity theory says that being in pain is a material, and Putnam's reply is that it looks much more like a job.

I want to give this failing its right name, because it is the one thing in this section that bears directly on machines. Carl Sagan, writing about the search for life elsewhere, coined the phrase "carbon chauvinism" for the assumption that life must be based on carbon because all the life we know is.⁶ The identity theory is carbon chauvinism about the mind. It may turn out to be true; it may be that only carbon can do the job. But that is something to be found out, not something to be built into the definition, and a theory that builds it in has begged the question this paper is asking.

Functionalism

The third position is the one that takes the chair seriously. Functionalism says that a mental state is defined by the job it does: by what typically causes it, what it typically causes in turn, and how it relates to the system's other states. Pain is the state that is typically caused by injury, that typically causes wincing and avoidance and the belief that something is wrong, and that interacts with other states in the ways pain does. Anything that has a state doing that job is in pain, whatever the state is made of.

Putnam arrived at this view by way of a comparison that is the ancestor of the whole present debate. In "Minds and Machines" he pointed out that a Turing machine, the abstract computing device Turing described, has two quite different descriptions.⁷ It has a physical description, in terms of its parts and what they are made of, and it has a logical description, in terms of its states and the rules that take it from one state to the next. The logical description says nothing about the parts. The same machine, in the logical sense, could be built out of relays or valves or transistors or, Putnam says, out of people passing slips of paper, and it would be the same machine because it would run through the same states in the same order. Putnam's suggestion was that the relation between a mind and a brain is like the relation between the logical machine and the physical one. The mind-body problem, on this view, is not a problem about two kinds of stuff. It is a problem about two kinds of description of one thing, and Putnam thought that once this was seen, the problem would look much less deep.

He later put the proposal directly: pain is not a brain state but "a functional state of a whole organism," and he offered this as a hypothesis, to be tested against the alternatives, rather than as a piece of analysis.⁸ Fodor has given the clearest statement of the idea for a general reader, using the example of a soft-drink machine.⁹ The machine that takes a dime for a drink has two internal states: one in which it has received nothing, and one in which it has received a nickel. Give it a nickel in the first state and it goes to the second; give it a nickel in the second and it delivers a drink and returns to the first; give it a dime in the first and it delivers at once. Fodor's point is that those two states are fully defined by what they do, by what takes the machine into them and out of them, and that any machine that goes through them in that order is the same machine for the purpose of selling drinks, whether it is made of steel or cardboard or is a person in a booth. Functionalism says mental states are like the states of the drink machine, only enormously more numerous and more richly connected.

I find this the most convincing of the positions, and I want to say why in terms of something I have done rather than something I have read. This year I have written programs in two languages on the department's computers, and I have run them on more than one machine. When a program runs on a second computer it gives the same answers as on the first, and nobody says it has become a different program. The program is defined by what it

does with its inputs and how its states follow one another, and the computer is only what it happens to be running on. Functionalism is the claim that a mind is to a brain as a program is to a computer, and I think that claim fits every fact Section 2 established. It explains why behavior is the criterion, since behavior is what the functional states produce. It explains why the super-Spartan is in pain, since he has the state even though he has suppressed one of its usual effects. And it explains why the octopus can be in pain, since it has a state doing pain's job in a different material. Functionalism is, so to speak, the philosophy of the wooden clock.

Functionalism also gives the question of this paper its exact form. If mental states are functional states, then the question whether an inorganic system can be conscious is the question whether an inorganic system can have states that do the jobs that conscious states do. It is not a question about silicon or carbon. It is a question about organization, and it is answered, if at all, by looking at what a system does and how, not by looking at what it is made of. That is the form in which I will pursue it from here on.

Two things functionalism does not settle

I do not want to claim more for the position than it can bear, and there are two objections that I will name now and answer later, because they need the arguments of later sections.

The first is Ned Block's.¹⁰ Block asks us to imagine the entire population of China, each person equipped with a two-way radio, arranged to send each other signals in exactly the pattern in which the neurons of a human brain send each other signals, so that the whole population for an hour runs through the functional states of a human mind. Functionalism says that this system, for that hour, has a mind: it believes things, wants things, perhaps feels pain. Block's objection is that this is absurd, that there is plainly nobody home, and that a theory which says otherwise has left out whatever it is that makes a mind a mind. I take this objection seriously and I will not answer it here. I will only note that the feeling of absurdity is the same feeling Descartes had about his automata, and that Section 2 gave some reason to distrust that feeling. The full reply belongs to Section 6, after Hofstadter has said what a self is.

The second objection is the one I flagged at the end of Section 2 and will take up in Section 7. Functionalism tells us what pain does. It does not tell us why doing that should feel like anything. Suppose a system has a state that is caused by injury, causes avoidance, and connects to its other states just as pain does in me; the question remains whether there is something it is like for that system to be in the state, and functionalism, as I have stated it, does not answer. This is the inside, the thing Ryle and Wittgenstein left standing, and it has not gone away. I want to be honest that it is a real gap. What I will argue in Section 7 is

that it is a gap in every theory, including the identity theory, and that it is a gap in our knowledge of other people as well, so that it cannot be used against machines in particular without being used against everyone.

Eliminative materialism, briefly

For completeness I should mention the fourth position on Churchland's map, which is his own. Eliminative materialism holds that our ordinary talk of beliefs, desires and pains is a theory, an old and bad one, which a mature neuroscience will not reduce to brain states but simply replace, as chemistry replaced talk of phlogiston rather than finding out what phlogiston was made of.¹¹ I think this is a possibility that cannot be ruled out, but I do not think it changes the question of this paper. If our ordinary notion of consciousness is one day replaced by a better one, the question whether an inorganic system can have whatever the better notion picks out will still be there to ask. It will just be asked in new words.

The working hypothesis

Let me state where the paper now stands. Dualism is closed. Behaviorism confuses the criterion with the thing. The identity theory is carbon chauvinism, and begs the question. Functionalism defines mental states by their jobs, fits the facts of Section 2, and turns the machine question into a question about organization. I adopt it as the working hypothesis, with two debts to pay: Block's nation of China, which waits for Section 6, and the inside, what it is like to be the system, which waits for Section 7. The next two sections take up the two tests Descartes proposed, beginning with the test of language, which Turing turned into a game.

Notes to Section 3

1. Paul M. Churchland, *Matter and Consciousness: A Contemporary Introduction to the Philosophy of Mind* (Cambridge, Mass.: MIT Press, 1984), chapter 2, "The Ontological Problem (the Mind-Body Problem)," pp. 7-49.
2. Hilary Putnam, "Brains and Behavior" (1963), in *Mind, Language and Reality: Philosophical Papers*, vol. 2 (Cambridge: Cambridge University Press, 1975), pp. 325-341, especially p. 331 on the super-Spartans.
3. U. T. Place, "Is Consciousness a Brain Process?", *British Journal of Psychology* 47 (1956), pp. 44-50; J. J. C. Smart, "Sensations and Brain Processes," *Philosophical Review* 68 (1959), pp. 141-156; D. M. Armstrong, *A Materialist Theory of the Mind* (London: Routledge and Kegan Paul, 1968).
4. Smart, "Sensations and Brain Processes," p. 145.
5. Hilary Putnam, "The Nature of Mental States" (1967), in *Mind, Language and Reality*, pp. 429-440. The octopus example is Putnam's; the extraterrestrial is a natural extension of it.
6. Carl Sagan, *The Cosmic Connection: An Extraterrestrial Perspective* (New York: Doubleday, 1973), p. 42.
7. Hilary Putnam, "Minds and Machines," in Sidney Hook, ed., *Dimensions of Mind* (New York: New York University Press, 1960); reprinted in *Mind, Language and Reality*, pp. 362-385.
8. Putnam, "The Nature of Mental States," p. 433.

9. Jerry A. Fodor, "The Mind-Body Problem," *Scientific American* 244 (January 1981), pp. 114-123. The soft-drink machine is Fodor's example; I have simplified it slightly.
10. Ned Block, "Troubles with Functionalism," in C. Wade Savage, ed., *Perception and Cognition: Issues in the Foundations of Psychology*, Minnesota Studies in the Philosophy of Science, vol. 9 (Minneapolis: University of Minnesota Press, 1978), pp. 261-325, especially pp. 278-281 on the Chinese-nation example.
11. Churchland, *Matter and Consciousness*, pp. 43-49; see also Paul M. Churchland, "Eliminative Materialism and the Propositional Attitudes," *Journal of Philosophy* 78 (1981), pp. 67-90.

4. The Test of Language: Descartes Reversed

Descartes offered two tests for telling a machine from a man, and the first was language. In this section I want to follow that test through three hundred years, because something odd happened to it on the way. Descartes proposed it as a test the machine would certainly fail. Turing took the same test and proposed it as one the machine might pass, and said that passing would settle the matter. Between them stand the machines we have actually built, one of which passed the test with a good many people before anyone meant it to, and a philosopher from my own university who argued that passing the test settles nothing. My conclusion will be in the middle: the test is evidence, and it is the best evidence we could have, but it is not a proof, and I will try to say why that is not the disappointment it sounds like.

Descartes' first test

Here is the test as Descartes states it. Even if a machine were built to resemble us and imitate our actions, it "could never use words or other signs arranging them as we do in order to declare our thoughts to others."¹ He is careful about what he is and is not claiming. A machine could certainly be made to utter words, he says, and even to utter them in response to what is done to it: "if it is touched in a particular part it may ask what we wish to say to it; if in another part it may exclaim that it is being hurt, and so on." What it could never do is "arrange its speech in various ways, in order to reply appropriately to everything that may be said in its presence, as even the lowest type of man can do."²

By this he means that the difference between a man and a machine is not that the man makes sounds and the machine does not. It is that the man's sounds are answers. Say anything to him, anything at all, and he will reply to what you said; the machine has a fixed set of responses wired to a fixed set of triggers, and once you have found the edge of the set, there is nothing beyond it. Descartes' machine is like a music box: it plays its tune when the lid is opened, and it plays the same tune whether you have opened the lid to listen or by accident. A man asked the same question twice may give two different answers, and the second will take account of his having been asked before.

I think this is a good test and I do not want to weaken it. Its strength is that it does not depend on any theory of what thinking is. It asks only whether the thing can hold up its end of a conversation, and everyone knows what that means, whatever they believe about minds. Its weakness, which Descartes did not see because he could not have, is that he was sure of the answer in advance. He knew what machines were, in his century, and he knew they could not do it. He did not consider that the machines might change.

Turing's game

Turing considered exactly that. His paper "Computing Machinery and Intelligence" opens with the words, "I propose to consider the question, 'Can machines think?'," and then declines to answer it, on the ground that the words "machine" and "think" are too vague to support an answer.³ Instead he proposes to replace the question with another, "closely related to it and expressed in relatively unambiguous words," and the replacement is a game.

The game has three players: a man, a woman, and an interrogator who is in a separate room and communicates with the other two only by typed messages. The interrogator's job is to work out which of the two is the woman; the man's job is to make him guess wrong; the woman's job is to help him guess right. Now, Turing says, replace the man with a machine. Will the interrogator guess wrong as often as he did when a man was playing the part? That question, Turing proposes, is what we should ask instead of "Can machines think?"

Notice what this is. It is Descartes' test with the sign changed. Descartes said: a machine could never reply appropriately to everything said in its presence. Turing said: let us put the machine where it must reply to everything said in its presence, and see. The examination is the same. What has changed is that Turing does not know the answer beforehand, and he does not think anyone does. He makes a prediction, that within about fifty years machines will play the game well enough that an average interrogator will have no better than a seventy percent chance of guessing right after five minutes of questioning, but he offers it as a guess about engineering, not as a conclusion of philosophy.⁴

The second thing to notice is that Turing does not say that a machine which wins the game thinks. He is more careful than most people who quote him. What he says is that the original question is "too meaningless to deserve discussion," and that by the end of the century the use of words will have changed so much that one will be able to speak of machines thinking "without expecting to be contradicted."⁵ By this I take him to mean that "thinking" is a word whose use is fixed by what people are willing to apply it to, that up to now they have applied it only to things that can hold a conversation, and that when a machine can hold one, they will apply it to the machine. This is Wittgenstein's point from Section 2, put by an engineer: "look at the word 'to think' as a tool." Turing is not making a discovery about thinking. He is making a forecast about a word.

Turing's objections

Turing spends more than half of his paper answering objections, and two of them matter here. The first he calls the argument from consciousness, and he quotes it from a surgeon named Jefferson: "Not until a machine can write a sonnet or compose a concerto because of thoughts and emotions felt, and not by the chance fall of symbols, could we agree that

machine equals brain."⁶ By this Jefferson means what Descartes meant and what Block will mean in Section 6: that doing the thing is not enough, that there must be somebody home who feels it.

Turing's reply is short and I think it is the best thing in the paper. If the only way to be sure a machine thinks is to be the machine and feel oneself thinking, then the only way to be sure a man thinks is to be that man, and we are in the position of the solipsist, who can be sure of no mind but his own. "Instead of arguing continually over this point," Turing says, "it is usual to have the polite convention that everyone thinks."⁷ By this he means that we do not, in practice, demand proof that other people are conscious. We extend them the convention, on the evidence of what they say and do, and we would think a man who withheld it from his neighbors was mad. Turing's question is why the convention should stop at the edge of the species. He does not say it should not. He says that anyone who wants it to stop there owes an argument, and that the argument from consciousness is not one, since it proves too much.

The second objection Turing takes from Lady Lovelace, who wrote of Babbage's engine that it "has no pretensions to originate anything. It can do whatever we know how to order it to perform."⁸ This is Descartes' second test in a new dress, and I will hold it for Section 8. I will only note here that Turing's reply is that machines surprise him all the time, and that the surprise is not a sign of his carelessness but of the fact that a machine, once set going, does what its organization does and not what its designer had in mind. I said the same thing in Section 1 about Sartre's paper-knife, and I think Turing's experience of being surprised by his own programs is the best evidence there is that the designer's idea does not exhaust the machine.

A machine that passed

Now I want to look at what actually happened when a machine was put to Descartes' test, because it did not go as either Descartes or Turing expected. In the middle of the last decade Joseph Weizenbaum wrote a program called ELIZA, and gave it a script called DOCTOR, which made it imitate a psychotherapist of the kind that mostly reflects the patient's remarks back at him.⁹ The program is simple. It looks for key words in what is typed to it, and it has a small set of patterns for turning a remark into a question: tell it "My mother hates me" and it will say "Tell me more about your family." When it finds no key word it falls back on "Please go on" or on something you said earlier. A program of this kind is a standard exercise now, and I have seen how little there is inside one.

Weizenbaum was shocked by what happened next. His secretary, who had watched him write the program and knew exactly what it was, asked him to leave the room so that she could talk to it in private. Practicing psychiatrists proposed that it be put to clinical use.

People who conversed with it for a few minutes were sure that it understood them.¹⁰ Here was Descartes' machine, built at last: touch it in one place and it asks what you want, touch it in another and it says you have hurt it, and nothing beyond. And people believed it.

I think the lesson of ELIZA is about the examiner, not the machine. Descartes' test assumes an examiner who is trying to find the edge of the machine's responses, and ELIZA's edge is easy to find; ask it a direct question about anything, and it changes the subject. But the people who talked to it were not trying to find the edge. They were trying to be understood, and a thing that reflected their words back at them was enough. This does not show that the test is bad. It shows that the test is only as good as the examiner, and that we are very willing examiners. We extend Turing's polite convention to anything that answers, and we extend it faster than Descartes imagined. I will come back to this at the end of the section, because it cuts both ways.

Gunderson's correction

The strongest argument that the test proves nothing even when it is passed properly comes from Keith Gunderson, in a paper written a few years after Turing's and developed in his *Mentality and Machines*.¹¹ Gunderson asks us to imagine a different game. The interrogator puts his foot through a hole in a wall. On the other side is either a person, who steps on his toes, or a box that drops rocks on them. If the interrogator cannot tell which is which, do we conclude that the rock box can step on toes? Gunderson's point is that we do not. We conclude that the rock box can play the toe-stepping game, which is to say that it can produce, by its own means, an effect that a foot also produces. Nothing follows about the box having feet.

Applied to Turing's game, the argument runs like this. A machine that wins the imitation game has shown that it can produce, by its own means, the effect that a thinking person also produces, namely appropriate replies. Nothing follows about the machine thinking, any more than feet follow from the rocks. Gunderson's later way of putting it is that the game tests what he calls program-receptive features of mind, the ones that can be specified as a task and so can be programmed, and leaves untouched what he calls program-resistant features, which are the ones Jefferson was pointing at: the feelings, the something it is like.¹²

I think Gunderson is right, and I want to say precisely what he is right about, because it is less than is sometimes claimed. He has shown that passing the test is not the same as thinking, and that no amount of passing adds up to a proof. He has not shown that passing is not evidence. The rock box is a poor analogy in one respect: we know exactly how the rocks are dropped, and knowing that, we know there are no feet. In Turing's game we are not given the mechanism; we are given only the conversation, which is exactly what we are given with other people. If the interrogator in the toe-stepping game were forbidden to look

behind the wall, then a partner who stepped on his toes in reply to everything he did, faster when he was slow and gently when he was tired, would give him good reason to think there was a foot back there. He would not have a proof. He would have what he has about everyone.

The verdict

Let me put the results together. Descartes' first test asks whether a thing can reply to whatever is said to it. Turing showed that the test can be given without a theory of thinking and proposed that passing it should change how we use the word. ELIZA showed that we give the test badly, because we want to be understood more than we want to find the edge. Gunderson showed that passing it, even when it is given well, is not a proof.

My verdict is this. The test is evidence and not criterion. A machine that replied appropriately to everything said in its presence would have given us the strongest evidence we can have of a mind, because it would have given us the same evidence that other people give, and we do not ask more of them. It would not have given us a proof, and there is no proof to be had, for the machine or for the neighbors. I accept that asymmetry, and I do not think it is a defect in the test. It is the position we have always been in. What Gunderson has done is remind us that we are in it, and what ELIZA has done is warn us that we forget it easily.

There is one more thing to say, and it is the reason I have spent so long on a test I do not think settles anything. Someone who refuses the evidence in the machine's case while accepting it in his neighbor's owes an account of the difference, and the only difference he can point to is what the two are made of. That is the identity theory again, and Section 3 has already said what is wrong with it. So the test, which proves nothing, does something after all. It forces the carbon chauvinist into the open, where he has to say that a conversation counts as evidence of a mind when it comes from meat and not when it comes from metal, and then to say why. Section 5 takes up the best attempt anyone has made to say why.

Notes to Section 4

1. René Descartes, *Discourse on the Method*, Part V, in *The Philosophical Works of Descartes*, trans. Elizabeth S. Haldane and G. R. T. Ross, vol. 1 (Cambridge: Cambridge University Press, 1911), p. 116.
2. *Ibid.*, p. 116.
3. A. M. Turing, "Computing Machinery and Intelligence," *Mind* 59 (1950), p. 433; reprinted in Douglas R. Hofstadter and Daniel C. Dennett, eds., *The Mind's I* (Harmondsworth: Penguin, 1982), pp. 53-67.
4. *Ibid.*, p. 442.
5. *Ibid.*, p. 442.

6. Ibid., p. 445, quoting G. Jefferson, "The Mind of Mechanical Man," Lister Oration for 1949, *British Medical Journal* 1 (1949), pp. 1105-1121.
7. Turing, p. 446.
8. Turing, p. 450, quoting Lady Lovelace's notes to her translation of Menabrea's memoir on the Analytical Engine (1842).
9. Joseph Weizenbaum, *Computer Power and Human Reason: From Judgment to Calculation* (San Francisco: W. H. Freeman, 1976), chapter 1. The program was first described in Weizenbaum, "ELIZA: A Computer Program for the Study of Natural Language Communication Between Man and Machine," *Communications of the ACM* 9 (1966), pp. 36-45.
10. Weizenbaum, *Computer Power and Human Reason*, pp. 5-7.
11. Keith Gunderson, "The Imitation Game," *Mind* 73 (1964), pp. 234-245; reprinted in Alan Ross Anderson, ed., *Minds and Machines* (Englewood Cliffs, N.J.: Prentice-Hall, 1964). Gunderson taught at the University of Minnesota, and I first met the argument there.
12. Keith Gunderson, *Mentality and Machines*, 2nd ed. (Minneapolis: University of Minnesota Press, 1985). The toe-stepping game is in chapter 2, "The Imitation Game," pp. 43-44; the distinction between program-receptive and program-resistant features is introduced in chapter 3, "Robots, Consciousness, and Programmed Behavior," pp. 74-75, and developed in chapter 5, "Some Mental Limitations of Some Machines," pp. 147-160. In the postscript to this edition Gunderson compares the toe-stepping game with Searle's Chinese room, pp. 197-204.

5. The Chinese Room

Section 4 ended by asking the carbon chauvinist to explain why a conversation counts as evidence of a mind when it comes from a person and not when it comes from a machine. John Searle has given the best answer anyone has, and it is the argument this paper most needs to meet. Searle does not deny that a machine could pass Turing's test. He grants it, and then argues that passing shows nothing, because the machine would be doing something that is not understanding and could never become understanding, however well it was done. In this section I will state his argument at full strength, set out the replies that were made to it when it appeared, and then give the reply I think is right, which turns on a distinction between levels of description that I take from Hofstadter. Before any of that I need to do what I promised in Section 1 and say what "understand" means, because I think Searle's argument depends on the word doing two jobs at once.

Three senses of "understand"

When we say someone understands Chinese, we may mean any of three things, and they come apart.

The first is that he can use the language: he can answer questions put to him in it, follow instructions given in it, and say in it what he wants to say. This is understanding as an ability, and it is the sense in which a school examiner certifies that a pupil understands French. Nothing about the inside of the pupil is examined; what is examined is what he can do.

The second is that his words are connected to the world: that when he says the Chinese word for "rain," he means rain, the wet stuff, and not merely a sound that goes with other sounds. This is understanding as meaning, or, as philosophers say, semantics, and it is the sense in which a parrot that can say "it is raining" does not understand what it says, since nothing in the parrot links the noise to the weather.

The third is that there is something it is like for him to understand: a click, a feeling of the words making sense, the experience Wittgenstein describes when a pupil working through a series suddenly says "Now I can go on!"¹ This is understanding as an experience, and it is the sense in which one can say, of a sentence that one has parsed correctly and can even translate, "I still do not really understand it."

Wittgenstein spent a good many sections of the *Investigations* on the third sense, and his conclusion was that it is a mistake to think understanding is the feeling.² The feeling of "Now I can go on" can occur and be followed by failure, and one can go on perfectly well without any feeling at all. What makes it true that the pupil understands is that he goes on

correctly; the click, when it comes, is a sign, and an unreliable one. I think this is right, and I want to keep it in view, because Searle's argument gets its force from the third sense while its conclusion is stated in the second.

Searle's argument

Here is the argument as Searle first gave it.³ Imagine that Searle, who knows no Chinese, is locked in a room. Through a slot he is passed batches of paper covered with Chinese characters. He has a book of rules, in English, which tells him that when he receives characters of such and such shapes he is to look up certain other characters and pass them back out through the slot. The rules are very good; they have been written by people who know Chinese and know how to answer questions in it. The result is that the people outside the room, who are passing in questions and receiving answers, cannot tell the room from a native speaker of Chinese. The room passes Turing's test. Yet Searle, inside it, understands nothing. He is shuffling symbols according to their shapes. He does not know what any of them mean, and no amount of shuffling will teach him.

Now, Searle says, the room is exactly what a computer is. A computer program is a set of rules for manipulating symbols according to their shapes, which is to say their syntax. Searle in the room is running a program, and running it perfectly. If running the program were sufficient for understanding Chinese, Searle would understand Chinese. He does not. Therefore running a program is not sufficient for understanding, and no program, however good, could give a machine a mind.

In the Reith Lectures Searle set the argument out as four premises and a conclusion, and I will give them in his order because the structure matters.⁴ First: brains cause minds. Second: syntax is not sufficient for semantics. Third: computer programs are entirely defined by their formal, or syntactical, structure. Fourth: minds have mental contents, specifically semantic contents. From the second, third and fourth it follows that no program by itself is sufficient to give a system a mind. From this and the first it follows that the way brains cause minds cannot be solely by running a program, and that anything else that caused a mind would have to have "causal powers at least equivalent to those of the brain."

I want to say at once that I think the argument is valid, and that its second premise is true. Syntax is not sufficient for semantics. A book of rules for shuffling shapes does not, by itself, connect any shape to the world, and a man following such a book understands nothing in the second sense of the word. Where I disagree with Searle is over what the argument shows about machines, and to get at that I need to look at the replies he considered, and at why he thought they failed.

The replies

Searle's paper was published with commentary from a couple of dozen critics, and he grouped their replies under several headings.⁵

The first he called the systems reply. The man in the room does not understand Chinese, the reply says, but the man is not the whole system. The system is the man together with the rule book, the paper, the pencils and the room, and it is the system that understands. Searle's answer is to have the man memorize the rules and do all the work in his head. Now the man is the whole system, and he still understands nothing.

The second he called the robot reply. Put the program in a robot, the reply says, with cameras for eyes and arms that move, so that its symbols are connected to the world by perception and action; then the symbols will have meaning. Searle's answer is that this changes nothing, because the man in the room, now receiving symbols from the cameras and sending symbols to the arms, is still shuffling shapes and still knows nothing about what the cameras see.

The third he called the brain simulator reply. Suppose the program simulates, neuron by neuron, the brain of a Chinese speaker, so that the man in the room is operating a system of water pipes and valves that reproduces exactly what the speaker's neurons do. Searle's answer is that the man turning the valves understands no Chinese, and neither do the pipes.

Searle also has a general answer, which he gives more than once and which I think is the real heart of his position. A simulation of a thing is not the thing. "No one supposes that computer simulations of a five-alarm fire will burn the neighborhood down," he writes, "or that a computer simulation of a rainstorm will leave us all drenched."⁶ Why, then, should a simulation of understanding understand? By this he means that understanding is a real effect, like fire and wetness, produced by real causes, and that the causes are, so far as we know, in the biology of brains. Brains produce minds, on his view, the way stomachs produce digestion: by their specific chemistry, and not by any pattern that could be lifted off them and run elsewhere. This is the biological account of mind at issue here.

The reply from levels

I want to give a reply that is a version of the systems reply, but I want to give it in a way that shows why Searle's answer to that reply does not work, and for this I need Hofstadter.

In *Gödel, Escher, Bach* there is a dialogue in which an anteater describes his friend Aunt Hillary, who is an ant colony.⁷ Aunt Hillary can hold a conversation; the anteater carries on long talks with her. No ant in the colony can hold a conversation, or knows that one is going on. Each ant follows a few simple rules about which way to go and what to do when it meets another ant. Out of millions of ants following such rules there arise larger patterns, trails and teams and what Hofstadter calls signals, and out of the signals arise larger

patterns still, and at some level the patterns are symbols, and Aunt Hillary thinks with them. Ask an ant whether it understands the conversation and the question has no sense. Ask Aunt Hillary and she will answer you.

Hofstadter's point is that a system can have properties at one level of description that none of its parts have at a lower level, and that this is not a trick or a manner of speaking but the ordinary way complex things are. A single brick does not keep out the rain; a house does, and there is nothing in the house except bricks and what holds them together. A single neuron in Searle's head does not understand English. If it did, we would not need the other ten billion. Searle understands English, and he is nothing but neurons and what holds them together. To ask whether a neuron understands is to make Ryle's category mistake in the other direction: instead of looking for the University among the colleges, one is asking each college whether it is the University.

Now look again at the room. Searle asks us to notice that the man in the room does not understand Chinese, and he is right; the man does not. But the man is playing the part of a neuron, or rather of the whole mechanism of neurons, the part of the thing that pushes the symbols around. He is the hardware. The understanding, if there is any, is not his to have or to lack, any more than the anteater's conversation belongs to any ant. It belongs, if it belongs to anything, to the system of rules and symbols that the man is executing, which is at a different level from him. Searle's intuition that there is no understanding in the room is an intuition about the man, and it is correct, and it is beside the point.

Searle's answer to the systems reply was to have the man memorize the rules, so that there would be no system but the man. I do not think this helps him, and the reason is instructive. When the man has memorized the rules and runs them in his head, there are now two systems in one skull: the man, who speaks English and knows nothing of Chinese, and the Chinese-speaking system, which he is running the way a computer runs a program. The man's testimony that he does not understand Chinese is the English system reporting on itself. It is not a report on the Chinese system, to which the man has no access, because it is not a part of his English-speaking mind; it is something his brain is doing alongside that mind. This sounds strange, and Searle counts on its sounding strange. But it is only what happens whenever one system runs another. The computer on which I run my programs has an operating system that knows nothing of what my programs are about, and a report from the operating system that it does not understand Prolog would be true and would tell us nothing about the Prolog program.

The same reply meets the brain simulator case, and here it turns against Searle directly. If the man's water pipes reproduce, valve for valve, what the neurons of a Chinese speaker do, then either the pipes understand Chinese at the level of the whole system, or something in the speaker's brain besides the pattern of what the neurons do is needed for understanding. Searle takes the second horn; that is the biological position at issue. But

notice what the second horn commits him to. It commits him to saying that what produces understanding is not the organization of the brain but its material, its particular chemistry, so that the same organization in pipes would not do. That is the identity theory, and Section 3 has already said what is wrong with it. It is carbon chauvinism, and the Chinese room does not argue for it. The Chinese room shows, at most, that the man does not understand; the step from there to "and so the chemistry must be doing it" is an assumption, not an inference.

What is left of Searle

I said I would try to be fair to the side I reject, and I want to end this section by saying what I think Searle has actually shown, because it is not nothing.

He has shown that syntax is not sufficient for semantics, and I accept this. A system whose symbols are connected to nothing outside itself does not understand in the second sense, whatever it says. This is why the robot reply matters more than Searle allows. A system whose symbol for rain is connected, by the way the system works, to the cameras that see the rain and the arms that reach for the umbrella has something that the man in the room lacks, and it is exactly the thing the second sense of "understand" asks for. Searle's answer, that the man in the robot still shuffles shapes, is the level confusion again: of course the man does, and of course the neuron does.

He has shown, too, that a simulation is not the thing simulated, and this is true where the thing is fire or rain. But it is not obviously true where the thing is a pattern. A simulation of a calculation is a calculation; a simulation of a game of chess is a game of chess. The question is which kind of thing understanding is, and Searle's examples assume the answer. If understanding is a matter of what a system does and how its parts are organized, then a simulation of it, done in enough detail, is another instance of it, as a simulation of chess is another game. If understanding is a matter of what the parts are made of, then a simulation is only a picture. Section 3 gave reasons for the first answer. Searle has given none for the second, beyond the intuition of the man in the room, and I have tried to show that the intuition is about the wrong level.

Finally, Searle is willing to say something that many of his admirers are not. Asked whether a machine could think, he answers: yes, we are precisely such machines.⁸ Brains are machines, on his view, and a machine made of something other than brain-stuff could think too, provided it had "causal powers at least equivalent to those of the brain." What he denies is only that running a program is enough. So the distance between Searle and this paper is smaller than the noise around the Chinese room suggests. We agree that a machine could be conscious. We agree that it would need the right causal powers. We disagree about whether we know those powers to be chemical rather than organizational, and I have argued that we do not, that the Chinese room does not show it, and that the position which

assumes it has a name and a known defect. Searle asserts the biological requirement. He does not show it. Until he does, the room stands as a very clear picture of a level confusion, and the door it was meant to shut is still open.

Notes to Section 5

1. Ludwig Wittgenstein, *Philosophical Investigations*, trans. G. E. M. Anscombe, 3rd ed. (Oxford: Basil Blackwell, 1967), §151.
2. Ibid., §§151-155 and §§179-184.
3. John R. Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3 (1980), pp. 417-424; reprinted in Douglas R. Hofstadter and Daniel C. Dennett, eds., *The Mind's I* (Harmondsworth: Penguin, 1982), pp. 353-373.
4. John R. Searle, *Minds, Brains and Science: The 1984 Reith Lectures* (London: British Broadcasting Corporation, 1984), lecture 2, "Can Computers Think?", pp. 39-41.
5. Searle, "Minds, Brains, and Programs," *The Mind's I*, pp. 358-367. The commentaries appear in *Behavioral and Brain Sciences* 3 (1980), pp. 424-457.
6. Ibid., *The Mind's I*, p. 370.
7. Douglas R. Hofstadter, *Gödel, Escher, Bach: An Eternal Golden Braid* (Harmondsworth: Penguin, 1980), "Prelude... Ant Fugue," pp. 275-284, and chapter X, "Levels of Description, and Computer Systems," pp. 285-336.
8. Searle, "Minds, Brains, and Programs," *The Mind's I*, p. 372; *Minds, Brains and Science*, pp. 40-41.

6. Strange Loops and the Self

Sections 4 and 5 were about tests and objections, and they cleared ground without building on it. In this section I want to build. The working description of consciousness in Section 1 had three parts, and the middle part, a point of view sustained by a model of the system itself, was the one I said came nearest to explaining how consciousness could arise at all. That part is Hofstadter's, and I now want to set it out properly, because it does something that none of the other positions in this paper do. It gives an account of where a self comes from, and the account, by its nature, says nothing about what the self is made of. After setting it out I will answer the one serious objection to it, which comes from Gödel by way of Lucas, and then I will use it to pay the debt to Block that has been outstanding since Section 3.

Self-reference as the mechanism

Hofstadter's book begins with Gödel's theorem, and the theorem is worth stating, because the whole account of the self is modelled on it.¹ Take a formal system rich enough to express arithmetic: a set of symbols, a set of axioms, and rules for deriving new strings of symbols from old ones. Gödel showed that inside any such system one can construct a string which, read as a statement about the system, says "this string cannot be derived in this system." If the system is consistent, the string cannot be derived, since deriving it would make the system assert a falsehood; but then what the string says is true. So every consistent system of that kind contains a truth it cannot prove.

What Hofstadter takes from this is not the limitation but the mechanism. The Gödel string is possible because the system is rich enough to talk about itself. Its symbols, which were meant to be about numbers, turn out to be able to be about strings of symbols, including the string they are in. Hofstadter calls this a strange loop: a situation in which, moving through the levels of a system, one finds oneself back where one started, the way one does on an Escher staircase or in the drawing of two hands each drawing the other.² The loop is not a defect. It is what happens whenever a system becomes powerful enough to represent the thing that is doing the representing.

Now apply this to a brain. A brain is a system of symbols, in the sense that it contains structures which stand for things in the world: for the coffee on the desk, for the rain outside, for the professor across the table. Hofstadter's proposal is that a brain rich enough to represent the world will, like Gödel's system, find itself able to represent itself, and that when it does, one of its symbols will be a symbol for the whole system of which it is a part.³ That symbol is the self. It is not a ghost and it is not a part of the brain that one could point to on a diagram. It is a pattern in the brain's activity which stands for the brain, in the way

that a map may have on it a small mark that stands for the map, with the words "you are here." A map without that mark is a picture of a place. A map with it is a picture of a place that includes the picture, and the person reading it now knows where he stands.

I think this is the best account of a point of view that anyone has given, and I want to say plainly what it achieves. Nagel said that a conscious thing is one for which there is something it is like to be it. A thing that models the world has a picture of the world. A thing that models itself within that picture has a place in the picture, a here from which everything else is there, and that is what a point of view is. The self-symbol does not add a second thing to the system. It adds a mark on the map, and the mark makes the map a view.

Hofstadter puts the general idea in a sentence that I want to quote, because it is the thesis of this section: "My belief is that the explanations of 'emergent' phenomena in our brains, for instance, ideas, hopes, images, analogies, and finally consciousness and free will, are based on a kind of Strange Loop, an interaction between levels in which the top level reaches back down towards the bottom level and influences it, while at the same time being itself determined by the bottom level."⁴ By this he means that the self, once it exists as a pattern, is not idle. The system's picture of itself changes what the system does, which changes the picture, and so on, without end. Descartes wanted the mind to move the body and could not say how a ghost could push. Hofstadter's answer is that the mind is a pattern in the body's activity, and a pattern in a thing's activity can affect that activity in the way a whirlpool, which is nothing but water moving, determines where the water goes.

Two things follow that matter for this paper. The first is that the account is neutral about materials from top to bottom. Gödel's theorem does not care what the symbols are written in. A self-model is a pattern of symbols standing for the system that carries them, and it can be carried by anything that can carry symbols. Nothing in the story requires neurons, and if one wants to insist on neurons one has to add that insistence from outside, which is the move Section 3 already found wanting. The second is that the account is a hypothesis. Hofstadter offers it as one, and so do I. It says how a point of view could arise in a physical system. It does not prove that this is how ours did, and it does not, as I will grant in Section 7, say why a point of view should be accompanied by anything it is like. But it is a hypothesis of the right shape, the only one I know of, and a machine built on it would be a machine with a self in the only sense of "self" I can make clear.

Lucas's objection

The account leans on Gödel, and there is a famous argument which uses Gödel to show the opposite of what I have just claimed. It is J. R. Lucas's, and it runs as follows.⁵ Take any machine, which is to say any formal system. Gödel's theorem gives us a sentence, its Gödel sentence, which is true but which the machine cannot prove. But we can see that the

sentence is true; that is what Gödel showed us how to do. So there is something we can do that the machine cannot, and since this holds for every machine, we are not machines. Minds are not formal systems, and no formal system can be a mind.

I think the argument is wrong, and Hofstadter's reply to it seems to me decisive, so I will give it briefly.⁶ The step that fails is "we can see that the sentence is true." What Gödel's theorem actually says is: if the system is consistent, then its Gödel sentence is true and unprovable. To see that the sentence is true, then, one must know that the system is consistent. For a simple system one may know that. For a system as complicated as a human mind, nobody knows it, and there is good reason to think it is false: people hold contradictory beliefs all the time and do not collapse. So the mathematician who "sees" the truth of a machine's Gödel sentence is relying on a premise about the machine, and when the machine in question is himself, he has no way of establishing the premise. Lucas has shown that a man can see the blind spot of a machine simpler than himself. He has not shown that the man has no blind spot. Everyone can see the back of someone else's head, and it would be a poor argument that concluded from this that one's own head has no back.

There is a second reply, which Hofstadter also gives, and which I find even more telling. Whatever the mathematician does when he "sees" the truth of the Gödel sentence, he does with his brain, which is a physical system following physical laws. If seeing the truth of Gödel sentences is a thing that a physical system can do, then it is a thing that a machine can be built to do, and Lucas's argument, applied to that machine, produces a new Gödel sentence which the new machine cannot prove, and so on. There is no level at which the mathematician escapes the regress, because he is in it too. Lucas has mistaken the fact that every formal system has a limit for the claim that minds have none. What Gödel showed is that there is no top to the hierarchy of systems. That is a fact about all of them, and it is the fact from which Hofstadter built the self.

The self is a pattern

I now want to bring in two arguments that come at the same conclusion from a quite different direction, because they show that the account of the self I have given is not only a theorist's convenience. It matches what we find when we ask, in plain cases, where a self is.

The first is Dennett's, in a piece called "Where Am I?"⁷ Dennett imagines that his brain has been removed and placed in a vat in Houston, and connected by radio to his body, which carries on as before, the nerves sending their signals to the vat and the vat sending its instructions back. Standing in the laboratory and looking at his brain in its vat, he asks where he is, and finds that the answer is not obvious. He is inclined to say he is where his body is, since that is where his point of view is; his brain is over there, in the tank. Later a computer is programmed to duplicate the brain's activity and run in parallel with it, and a

switch is installed so that either the brain or the computer can be in control of the body; and when the switch is thrown, Dennett notices nothing. Whichever is running him, he is where his body is, and he is himself.

By this Dennett means that the self does not sit in any particular lump of matter. It goes where the pattern goes. When the pattern is carried by the brain in the vat, he is that; when it is carried by the computer, he is that; and he has no way of telling, from the inside, which is which, because the inside is the same. This is not a proof that a computer could carry a self. It is a story, and Dennett tells it as one. But it is a story that makes vivid what the functionalist claim comes to, and it makes it vivid by asking the one person who ought to know where the self is, and getting the answer: with the pattern.

The second argument is Parfit's, and it comes at the same point through the question of personal identity.⁸ Parfit imagines a machine that records the exact state of every cell in a person's body, destroys the body, and transmits the record to Mars, where a replica is built, cell for cell, with all the same memories and intentions. Is the person on Mars the same person who stepped into the machine on Earth? Parfit's answer is that the question has no deep answer, and does not need one. What matters, he argues, is not identity but what he calls psychological continuity: that the person on Mars remembers what the person on Earth did, intends what he intended, and carries on his projects. If that continuity holds, then everything that mattered about the person has survived, whatever we say about "the same." The self, once again, is the pattern, and the matter is what the pattern happens to be in.

I find these two arguments convincing together in a way that neither is alone. Dennett shows that the self follows the pattern between a brain and a machine. Parfit shows that it follows the pattern between one body and another. Neither is about consciousness in machines directly. But if the self is a pattern, then the question of Section 1, whether consciousness needs living matter, becomes the question whether the pattern needs living matter, and nothing about a pattern says so. A tune is not the piano. The same tune can be played on a different piano, or on a flute, and it is a tune all the same, and if one asks whether it is really the same tune the answer is Parfit's: it has everything that mattered.

The nation of China

I can now pay the debt to Block.⁹ He asked us to imagine the population of China, linked by radios, running through the functional states of a human mind for an hour, and he said that functionalism must call this a mind and that this is absurd, because there is plainly nobody home.

Here is the reply. If the self is a pattern that a system carries when it models the world and itself within the world, then the nation of China, running that pattern, carries a self, and it carries it in exactly the sense in which a brain does. Ask any Chinese citizen whether

he understands what the nation is thinking and he will say no, and he will be right, and it will be beside the point, for the reason given in Section 5: he is an ant, and the question is about Aunt Hillary. The self that the nation carries, for that hour, is not any citizen's and not the sum of the citizens'. It is the pattern's, and there is nobody home in the citizens because the citizens are not where home is.

Why, then, does it feel absurd? I think for two reasons, and neither is an argument. The first is scale. Block's mind runs across a country and takes an hour to do what a brain does in an instant. We are not built to see a mind at that scale, any more than we are built to see a glacier flow. A glacier is a river, and it is moving, and if one filmed it for a year and ran the film in a minute one would see the river. Nobody watching a glacier for an afternoon sees anything but ice. Block has shown us the afternoon and asked whether we see the river. The second reason is the one I have named before. The feeling that there is nobody home is the feeling Descartes had when he looked at his automata, and Jefferson had when he thought of a machine writing a sonnet, and Searle had in his room. It is the feeling that where there is no ghost there is no one. Section 2 retired the ghost and Section 5 traced the feeling to a confusion of levels. It is a real feeling, and I share it, and I do not think it can be helped. But a feeling that has been wrong three times is not evidence the fourth time.

What Block's argument leaves standing is the same thing every argument in this paper has left standing, and I will name it once more before the section that is about it. The nation of China, running the pattern, would have a point of view in the sense I have made clear: a model of the world with a mark on it saying "here." Whether there would be something it is like to be the nation, whether the mark would be, so to speak, lit, is a question the account of the self does not answer. Block thinks it obvious that it would not be. I do not think it is obvious either way, and Section 7 is about why.

Notes to Section 6

1. Douglas R. Hofstadter, *Gödel, Escher, Bach: An Eternal Golden Braid* (Harmondsworth: Penguin, 1980), Introduction and chapter IX, "Mumon and Gödel." The formal construction is in chapter XIV, "On Formally Undecidable Propositions of TNT and Related Systems."
2. *Ibid.*, pp. 10-15, on the Escher prints and the notion of a Strange Loop.
3. *Ibid.*, chapter XX, "Strange Loops, Or Tangled Hierarchies," pp. 684-719, especially "Consciousness as an Intrinsically High-Level Phenomenon" and "Strange Loops as the Crux of Consciousness," pp. 708-709.
4. *Ibid.*, p. 709.
5. J. R. Lucas, "Minds, Machines and Gödel," *Philosophy* 36 (1961), pp. 112-127; reprinted in Alan Ross Anderson, ed., *Minds and Machines* (Englewood Cliffs, N.J.: Prentice-Hall, 1964).
6. Hofstadter, *Gödel, Escher, Bach*, chapter XV, "Jumping out of the System," especially pp. 471-477, and chapter XVII, "Church, Turing, Tarski, and Others," especially pp. 577-578, on Lucas.

7. Daniel C. Dennett, "Where Am I?", in *Brainstorms: Philosophical Essays on Mind and Psychology* (Hassocks: Harvester Press, 1979), pp. 310-323; reprinted in Douglas R. Hofstadter and Daniel C. Dennett, eds., *The Mind's I* (Harmondsworth: Penguin, 1982), pp. 217-229.
8. Derek Parfit, *Reasons and Persons* (Oxford: Clarendon Press, 1984), Part Three, chapter 10, "What We Believe Ourselves to Be," pp. 199-201, and chapter 12, "Why Our Identity Is Not What Matters."
9. Ned Block, "Troubles with Functionalism," in C. Wade Savage, ed., *Perception and Cognition*, Minnesota Studies in the Philosophy of Science, vol. 9 (Minneapolis: University of Minnesota Press, 1978), pp. 261-325, especially pp. 278-281. Compare Hofstadter's "A Conversation with Einstein's Brain," in *The Mind's I*, pp. 430-457, which asks the same question of a book containing a complete description of a brain.

7. The Intentional Stance and the Problem of Pain

Every section so far has ended by naming a debt, and it has always been the same one. Ryle and Wittgenstein removed the ghost and left something it is like to be in pain. Functionalism explained what pain does and did not say why doing it should feel like anything. The strange loop explained how a system comes to have a point of view and did not say whether the point of view is, as I put it, lit. This section is about that debt, and I want to say at the start that I will not pay it. Nobody has. What I will try to do is measure it accurately: to say what the functionalist account does cover, what it leaves out, and what the leaving out does and does not show about machines. I will take the two halves in turn, the outside of a mind and the inside, using Dennett for the first and Nagel for the second, and then I will say where that leaves the paper.

The outside: intentional systems

Dennett proposes that we distinguish three ways of predicting what a thing will do, which he calls stances.¹ From the physical stance we predict by the laws of physics and the thing's physical state: this is how we predict that a dropped stone will fall. From the design stance we predict by assuming that the thing will work as it was designed to: this is how we predict that an alarm clock will ring at seven, without knowing anything about its springs. And from the intentional stance we predict by assuming that the thing has beliefs and desires and will act rationally on them: this is how we predict what a chess player will do, whether the player is a man or a program.

Dennett's definition is this: "an intentional system is a system whose behavior can be (at least sometimes) explained and predicted by relying on ascriptions to the system of beliefs and desires."² By this he means that having beliefs is not a matter of containing some special inner item, a belief, that one could find on a dissecting table. It is a matter of being the kind of thing that the strategy works on. The chess program is an intentional system because, when I want to know what it will do next, the only practical way to find out is to ask what it wants (to protect its king) and what it believes (that my bishop threatens it). I could in principle predict its next move from the physical stance, by following the electrons, but I would die of old age first.

I want to notice what this does and does not claim, because Dennett is careful and his readers often are not. He does not claim that the chess program is conscious. He claims that the question whether something has beliefs is a question about which strategy works, and that the strategy works on the program. Whether the strategy is "really true" of the program,

whether the program "really" has beliefs as opposed to merely being usefully treated as if it does, is a question Dennett declines to answer, on the ground that it has no answer that could make a difference. The decision to adopt the stance, he says, is pragmatic.³

An example of my own may make the point and its limit clear together. We say the sky "threatens" rain, and nobody thinks the sky has intentions; the word is a figure of speech, and we could do without it. We say the dog "wants" to go out, and here we could not do without it: there is no way to predict the dog at the door except by ascribing the want, and the ascription works every time. The dog is an intentional system and the sky is not. But notice that this tells us nothing about whether there is something it is like to be the dog. It tells us that the strategy works. The strategy works on the dog, on the chess program, on the nation of China, and on my professor, and it works for the same reason in each case, which is that each is organized so that treating it as rational pays. That is what I meant in Section 1 by the third part of the working description, and it is the whole of the outside of a mind. It is what can be seen from the door.

Why you can't make a computer that feels pain

Dennett has also written directly about the inside, in a paper whose title states the objection this section has to meet: "Why You Can't Make a Computer That Feels Pain."⁴ The paper is not what its title suggests, and what it actually argues is the most useful thing I have found on the subject.

Dennett begins by trying to design a computer that feels pain. He does it seriously. He sketches a system with sensors for damage, thresholds, pathways that carry the signal to a center that assigns it priority, connections to the systems that control withdrawal and complaint and memory, and provision for the effects of morphine, of distraction, of the difference between a pain one attends to and a pain one has forgotten. The design gets very detailed, and Dennett's point is that at no stage does it run into a wall; every feature of pain that a physiologist could name can be put into the design.

Then he turns to the features that the physiologist does not name, the ones our ordinary concept of pain insists on, and he shows that they contradict each other. Our concept says that pain is essentially awful: a pain that did not hurt would not be a pain. But patients who have had certain operations, and patients on morphine, report that they still feel the pain and that it no longer bothers them; they have the sensation without the awfulness.⁵ Our concept says that a person cannot be wrong about whether he is in pain. But people are wrong about it in both directions, under anesthesia that leaves the behavior of pain and removes the memory, and under hypnosis. Our concept says that pain is a sensation, a thing that is felt; and it says that pain is a reaction, a thing that is done. Dennett's conclusion is that the ordinary concept of pain is not a coherent description of anything. It makes several demands at once, that pain be awful, that one cannot be mistaken about it, that it be a

sensation, that it be what makes one wince, and the clinical cases show that no single state, in any creature, meets all of them together. So if "feeling pain" means having a state that satisfies the whole concept, then nothing feels pain, people included. That is why, Dennett says, you cannot make a computer that feels pain: the specification contradicts itself, like an order for a round square, and the failure belongs to the concept and not to the machine.

I think this is right and I think it changes the shape of the question. The objection I have been carrying from section to section says: a machine might do everything pain does and still not feel it. Dennett's reply is that "feel it," in the ordinary sense, names a bundle that comes apart under examination, so that the objection is not a single objection but a family of them, and each member of the family is about some feature of pain that can either be put into the design or shown to be incoherent. What is left, when the family has been sorted, is not nothing. But it is a great deal smaller than "feeling," and it is time to say what it is.

The inside: Nagel's bat

What is left is Nagel's point, and it is best stated in his own words. "Every subjective phenomenon is essentially connected with a single point of view," he writes, "and it seems inevitable that an objective, physical theory will abandon that point of view."⁶ By this he means that a scientific account of anything is an account from no point of view in particular, one that any observer could check, and that what makes a mental state mental is that it is had from a point of view, so that the account leaves out the very thing it was meant to explain.

Nagel's example is the bat. A bat finds its way by sonar, and there is, Nagel assumes, something it is like to be a bat doing this. We can learn everything about how sonar works. We can imagine ourselves hanging upside down, flying at dusk, catching insects by echo. But that only tells us what it would be like for us to be a bat, and the question was what it is like for the bat. To that question all our knowledge of the bat's physiology gives no answer at all, and Nagel thinks it never could, because the answer is available only from the bat's point of view, and we are not there.

Jackson has put the same point as a story about a scientist, and I will give it because it is the strongest form of the argument.⁷ Mary has lived her whole life in a black and white room, and in it she has learned, from black and white books and screens, every physical fact there is about color and color vision: the wavelengths, the retina, the pathways, everything a completed science could tell. One day she is let out and sees a red tomato. Jackson says that at that moment she learns something, namely what it is like to see red, and that since she already knew every physical fact, what she learns is not a physical fact. So there are facts that are not physical, and materialism, in every form, including functionalism, is false.

I have thought about Mary a good deal and I am not sure the argument works, so I will say what I think and mark it as what I think. It seems to me that what Mary gains when she leaves the room may not be a new fact but a new ability, or a new acquaintance with an old fact. A man who has read every book about riding a bicycle learns something when he first rides one, but what he learns is not a fact about bicycles that the books left out; it is how to do it, and what it is like to do it, and both of those are things that only doing it can give. If this is right, then Mary's case shows that there are two ways of having a fact, from the outside and from the inside, and not that there are two kinds of fact.⁸ But I hold this loosely. The reply may only rename the problem; "what it is like to do it" is the very thing Nagel said no outside account could give, and calling it an ability does not explain why the ability comes with anything it is like. So I record the reply and do not lean on it.

Gunderson, whose distinction I used in Section 4, has a name for what is left, and it is the right name. The features of mind that can be specified as tasks he called program-receptive; they are the outside, the intentional stance, everything from the door. The rest he called program-resistant, and what resists is exactly Nagel's point of view: not what the system does, but that there is something it is like to be it doing it.⁹ I said in Section 4 that Gunderson was right and that he had shown less than is claimed. Here is the place to say what he has shown. He has shown that there is an inside, and that no test from the outside reaches it. That is true, and I accept it, and I now want to say what follows from it, because I think what follows is not what the carbon chauvinist needs.

What the gap shows

Here is the position of the paper as honestly as I can state it. Functionalism, with Hofstadter's account of the self, explains the structure of a mind: what it does, how it is organized to do it, how it comes to have a point of view. It is silent about whether the point of view is accompanied by anything it is like to have it. I have no answer to that question and I have found none. I do not know why there is something it is like to be me, and no theory I have read tells me.

But I want to make three observations about the silence, because they are what this paper turns on.

The first is that the silence is not functionalism's alone. The identity theory says that pain is the firing of C-fibres, and it does not say why the firing of C-fibres should feel like anything either. Searle says that brains cause minds by their biological powers, and he does not say why those powers should produce a point of view rather than merely producing behavior. Every theory of mind is silent on this point, and the silence is not a defect of any one of them. It is a gap in what we understand, and it is the same gap wherever one stands.

The second is that the silence is symmetrical. I do not know that there is something it is like to be my professor. I have the evidence of what he says and does, and I have Turing's polite convention, and that is all I have, and it is all anyone has about anyone. If the inside is an objection to machines, it is an objection to other people, and nobody takes it as one. The carbon chauvinist who says "but there might be nothing it is like to be the machine" is right, and there might be nothing it is like to be him, and he does not doubt it, and I do not doubt it, and we are both relying on the convention.

The third follows from the other two. What is left when the functional account is finished is a gap in explanation, what Levine has called the explanatory gap, and a gap in explanation is a limit on what we can know.¹⁰ It is not evidence about what is the case. Nagel's argument, if it works, shows that the point of view cannot be reached from outside; it does not show that only certain materials have one. Jackson's argument, if it works, shows that there are two ways of having a fact; it does not show that one of them requires carbon. Nothing in the gap says a word about what a system must be made of, and the chauvinist who uses it against machines is quietly adding the premise that Section 3 already rejected. He is saying that carbon lights the mark on the map and silicon does not, and he is saying it because the gap gives him a dark place to say it in. It does. But a dark place is not a reason.

I want to end this section with the only positive suggestion I have, and to mark it clearly as a suggestion. It may be that the question "why is the point of view lit?" is asking for something that does not exist, in the way the question "where is the University?" was. If Hofstadter is right that a self is a system's model of itself, then what it is like to be that system may simply be what having such a model is, from the only place a model can be had, which is inside it. On this view the lighting is not a further thing added to the map. It is what a map with a "you are here" on it is, for the one who is there. I cannot show this. It has the form of Ryle's answer to Descartes, and I find it attractive for that reason, but I know that it may only be the ghost's last hiding place given a new name. So I offer it, and I do not build on it, and I let the debt stand.

What the paper needs from this section is less than a solution. It needs only this: that the inside is real, that the gap between it and every account of it is the same for every theory and for every mind, and that the gap is a limit on knowing rather than a fact about matter. All three are true. The question of what a conscious thing must be made of is therefore not answered here, and not answered against machines. It remains what Section 3 made it, a question about organization, and Section 8 turns to the one test that bears on organization directly: the second of Descartes' two.

Notes to Section 7

1. Daniel C. Dennett, "Intentional Systems," in *Brainstorms: Philosophical Essays on Mind and Psychology* (Hassocks: Harvester Press, 1979), pp. 3-22.

2. Ibid., pp. 3-4.
3. Ibid., p. 7.
4. Daniel C. Dennett, "Why You Can't Make a Computer That Feels Pain," in *Brainstorms*, pp. 190-229.
5. Ibid., pp. 208-210, on morphine, prefrontal lobotomy, and the dissociation of the sensation of pain from its awfulness.
6. Thomas Nagel, "What Is It Like to Be a Bat?", *Philosophical Review* 83 (1974), p. 437; reprinted in Douglas R. Hofstadter and Daniel C. Dennett, eds., *The Mind's I* (Harmondsworth: Penguin, 1982), p. 393.
7. Frank Jackson, "Epiphenomenal Qualia," *Philosophical Quarterly* 32 (1982), pp. 127-136. Mary is on p. 130.
8. A reply of this kind has been suggested by Laurence Nemirow, in his review of Nagel's *Mortal Questions*, *Philosophical Review* 89 (1980), pp. 473-477, and by David Lewis, "Postscript to 'Mad Pain and Martian Pain'," in *Philosophical Papers*, vol. 1 (New York: Oxford University Press, 1983), pp. 130-132. I arrived at the bicycle before I found either, and I am not sure they would accept the way I have put it.
9. Keith Gunderson, *Mentality and Machines*, 2nd ed. (Minneapolis: University of Minnesota Press, 1985), chapter 5, "Some Mental Limitations of Some Machines," pp. 147-160, where the four criteria for the distinction are set out; see also pp. 74-75 and, in the postscript, pp. 221-225.
10. Joseph Levine, "Materialism and Qualia: The Explanatory Gap," *Pacific Philosophical Quarterly* 64 (1983), pp. 354-361.

8. The Test of Generality: The Universal Instrument

I said in Section 4 that Descartes' second test was the sharper of the two, and that I had left it until now because it is the one the machines we have actually fail. This section is about that failure. I want to state the test, show that the machines of the present kind fail it and say exactly why, look at the philosopher who has made the failure into an argument that no machine could ever pass, and then describe the change in the way machines are being built that is, I think, the first serious attempt to meet the test rather than to work around it. My conclusion will be that the second test, like the first, is a test of organization and not of material, and that its lesson is about rulebooks, not about silicon.

Descartes' second test

Here is the test in Descartes' words. Even if a machine did many things as well as we do or better, "they would without doubt fail in certain others, by which means we may discover that they did not act from knowledge, but only from the disposition of their organs. For while reason is a universal instrument which can serve for all contingencies, these organs have need of some special adaptation for every particular action. From this it follows that it is morally impossible that there should be sufficient diversity in any machine to allow it to act in all the events of life in the same way as our reason causes us to act."¹

By this he means something that the first test did not say. The first test was about language, and one might think a machine could fail it for a shallow reason, for want of a good enough vocabulary. The second test is about the shape of a mind. A mechanism, Descartes says, has one part for each thing it does. A clock has a part that moves the hour hand and a part that strikes the bell, and if you want it to do a third thing you must add a third part. Reason is not like that. Reason is one instrument that serves for everything, and the same reason that plans a journey can settle a quarrel or mend a fence, without a special part for each. So a machine, however many parts it has, will always meet a situation it has no part for, and there it will stop, and the stopping will show that it never understood the situations it handled either. It was only disposed to handle them.

An example of my own may fix the difference. A Swiss army knife has a blade, a saw, a screwdriver, a corkscrew, and a small pair of scissors, and it does each of those jobs tolerably. It has no tool for a job nobody foresaw, and when such a job comes it is only a lump of metal. A hand does every one of those jobs worse than the knife does, and it does any job at all. The hand is Descartes' universal instrument. The knife is his machine, and his claim is that no matter how many blades one adds, it never becomes a hand. I think this is

the best argument against the possibility this paper defends, better than the Chinese room, because it does not depend on an intuition about a man in a box. It depends on a fact about the machines, and the fact is true.

The machines we have

The machines that have made artificial intelligence a matter of public money are called expert systems, and they are Descartes' machines to the letter.² An expert system is a collection of rules of the form "if such and such is the case, then conclude so and so," gathered from a human expert by interview and written down for the machine to apply. The best known diagnoses infections from a patient's symptoms and laboratory results and recommends drugs, and within its field it does about as well as a specialist.³ Others find mineral deposits, configure computers for customers, and advise on tax. The theory behind them, in the words of the man who has done most to promote them, is that "knowledge is power": that what makes an expert an expert is not a general capacity for reasoning but a large stock of particular facts and particular rules, and that if one can get the stock into a machine the machine will be an expert too.⁴

I have spent this year learning to write programs of this general kind, in Prolog, which is a language built for exactly this purpose, and the experience has taught me two things. The first is that the theory works, within its field. A few hundred rules, well chosen, will answer questions that a layman cannot, and there is a real pleasure in watching the program reach a conclusion one did not see coming from rules one wrote oneself. The second is what happens at the edge of the field, and it is the thing Descartes predicted. Ask the medical program about a broken axle and it will do one of two things: it will say nothing, because no rule matches, or it will find some rule whose conditions the axle happens to satisfy and diagnose an infection. It does not know that the axle is not a patient. It does not know that there is a world beyond its rules, in which axles and patients are different kinds of thing. It is a knife with a great many blades and no hand, and its performance inside the field is no evidence that it understands the field, because the same performance is compatible with its understanding nothing, which is what the axle shows.

This failure has a technical name, or rather the deepest form of it does. It is called the frame problem, and Dennett has told a story that shows what it is.⁵ A robot is told that its spare battery is in a room, on a wagon, with a bomb on the wagon that is about to go off. The robot works out that pulling the wagon out of the room will bring the battery out, and it pulls the wagon, and the bomb comes too, and the robot is destroyed, because it never considered that pulling the wagon would move the bomb as well. So a second robot is built, which is told to consider the consequences of its actions. It sits in the room working out that pulling the wagon will not change the color of the walls, will not alter the number of

wheels on the wagon, will not affect the price of tea, and the bomb goes off. A third robot is built to ignore irrelevant consequences, and it sits in the room making an enormous list of the consequences that are irrelevant, and the bomb goes off.

Dennett's point is that a human being packing for a journey does not consider whether the color of his socks affects the flight, and does not have to consider it in order to ignore it. He simply does not think of it, and thinking of nearly everything and ignoring most of it is not what he does. Nobody knows how to give a machine that ability, and every attempt to give it by adding rules produces a machine that must consult the rules about which rules to consult, and then rules about that. This is the second test in its modern dress. The rulebook cannot contain its own application. Wittgenstein said this about rules in general, that no course of action could be determined by a rule, because every course of action can be made to accord with the rule on some interpretation,⁶ and the frame problem is that remark turned into an engineering difficulty. The machine that acts from the disposition of its organs meets an event of life that its organs were not disposed for, and there is no rule that tells it which of its rules to break.

Dreyfus

Hubert Dreyfus has made all of this into an argument that no machine could pass the second test, and his book has been the standing objection to the field for fifteen years.⁷ His case is that the whole enterprise rests on a mistake about what knowledge is. The field assumes that to know something is to have a description of it, a set of facts and rules that could be written down; Dreyfus argues, drawing on Wittgenstein and on Ryle's distinction between knowing how and knowing that,⁸ that most of what a person knows is not of that kind at all. The chess master does not consult rules; he sees the right move, and can often not say why. The person who walks into a room and knows something is wrong has not deduced it. Dreyfus's later work with his brother sets out five stages by which a beginner becomes an expert, and the striking thing about the progression is that the rules are what the beginner uses, and the expert has left them behind.⁹ The expert system, on this account, has caught the human expert at the wrong stage. It has written down what he says when asked to explain, which is the beginner's version, and it has left out what he actually does, which cannot be written down.

Dreyfus also argues that this know-how depends on having a body and being in a situation, so that a machine which is not in the world, which receives its facts by typewriter, could never have it. This is the second test again, with the body added as the reason for the universality: the hand can do anything because it is a hand, in a world, with things to handle.

I think Dreyfus is right about the machines we have and wrong about what follows. He has shown that intelligence is not a rulebook, and that the expert systems, which are rulebooks, are therefore not intelligent, and that adding rules will not make them so. That is a diagnosis of the failure of the second test, and it agrees with Descartes exactly: special adaptation for every particular action is not reason. But Dreyfus concludes that no machine could have the know-how the rulebook lacks, and here I think he has made the identity theorist's move in a new place. He has taken the fact that all the know-how we know of is in bodies, and turned it into a definition. The question is not whether a rulebook can be a hand. It cannot. The question is whether a hand must be made of flesh, and the failure of the rulebook does not bear on that at all. It bears on the rulebook.

The hope

I want to end this section with the change I mentioned, because it is the one development in the field that seems to me to take the second test seriously, and because it is where my own work this year has led me.

Turing saw the problem in the paper of Section 4, and his answer was not to write a bigger rulebook. It was to build what he called a child machine, a machine with as little built in as possible, and to educate it, so that whatever generality it came to have would be learned rather than installed.¹⁰ Nobody followed that advice for thirty years, because nobody knew how to make a machine learn anything worth knowing. That is now beginning to change, from two directions.

The first is Minsky's, in a book that has just appeared.¹¹ Minsky proposes that a mind is not one thing but a society of a great many small agents, each of them mindless, each doing one simple job, and that intelligence is the way the society is organized and not a property of any member. This is Hofstadter's ant colony from Section 5, offered as a design rather than a fable. Its bearing on the second test is this: a society has no central rulebook to run out. When a new situation arrives, no one agent has to have been disposed for it; it is enough that some combination of agents, pushed and pulled by the situation, produces a response, and Minsky's claim is that generality lives in the combinations and not in any part.

The second is the work on parallel distributed processing, which I mentioned in Section 1 and which goes further in the same direction.¹² Here the machine is built as a network of very simple units connected to each other by links of adjustable strength, on the pattern of neurons and synapses, and it is not programmed at all. It is shown examples, and a procedure adjusts the strengths of the links until the network's response to the examples is right, and then it is shown things it has not seen. I have read a report this year of a network that was shown a few thousand words of English with their pronunciations and learned, from that, to pronounce English text it had never seen, badly at first and then well, without

anyone having written a single rule about English spelling.¹³ Nobody could write those rules; English has no such rules, only a great many tendencies and exceptions. The network has know-how of the kind Dreyfus says cannot be written down, and it has it because nobody wrote it down. It was learned, and it lives in the strengths of the links, spread across the whole network, in no place in particular.

My own project this term has been a much smaller thing, and I do not want to claim anything for it except a lesson. It is a program, written in Lisp, which is given a handful of mathematical axioms and the rules for combining them, and is then left alone to make new statements out of the old ones and to keep the ones that turn out to be true.¹⁴ It does this. It has produced statements I did not put into it and had not thought of, and I will admit that watching it do so was the most enjoyable thing I have done with a computer. Lady Lovelace would say that it has originated nothing, since everything it found was already contained in the axioms, and she would be right. But Turing would say that it surprised me, and he would be right too, and the two remarks together are the lesson. The program lives entirely inside its rulebook. It will never meet an axle, because there are no axles in its world, and its generality is the generality of the axioms and not of anything beyond them; in Descartes' terms it is one blade of the knife, ground very fine. Yet what it does inside that world is not what I told it to do but what its organization does when set going, and the designer's idea did not exhaust it. That is a small thing. But it is the thing Section 1 said about Sartre's paper-knife, and it is the beginning of the difference between a machine that has a part for every action and a machine that finds its own actions among the parts it has.

I want to be careful about what this amounts to. It is a hope and not a result. The networks are tiny compared with a brain; nobody knows whether the methods that train a thousand units will train ten billion; and a machine that learns to pronounce English has not thereby learned anything about the frame problem. It is quite possible that the new approach will run into a wall of its own, as the last one did. What I am claiming is smaller. Descartes said that reason is a universal instrument and that a machine, having a special part for every action, could never have "sufficient diversity" to be one. The machines we have prove him right about them. The machines now being tried are built on the opposite principle: not a part for every action, but a great many parts for no action in particular, organized so that action arises from their combination and their organization can be changed by experience. That is a description of a hand, or of a brain, and whether it can be made to work is an open question. But it is a question about organization, and it is being asked, and the second test, which was Descartes' best reason for shutting the door, has turned into a specification for what would have to be on the other side of it.

Notes to Section 8

1. René Descartes, *Discourse on the Method*, Part V, in *The Philosophical Works of Descartes*, trans. Elizabeth S. Haldane and G. R. T. Ross, vol. 1 (Cambridge: Cambridge University Press, 1911), p. 116.
2. For the field's own account of itself see Margaret Boden, *Artificial Intelligence and Natural Man* (Hassocks: Harvester Press, 1977), and Aaron Sloman, *The Computer Revolution in Philosophy* (Hassocks: Harvester Press, 1978).
3. Edward H. Shortliffe, *Computer-Based Medical Consultations: MYCIN* (New York: Elsevier, 1976).
4. Edward A. Feigenbaum and Pamela McCorduck, *The Fifth Generation: Artificial Intelligence and Japan's Computer Challenge to the World* (Reading, Mass.: Addison-Wesley, 1983). "Knowledge is power, and the computer is an amplifier of that power" is the book's opening sentence, p. 1; the phrase is also the title of its second chapter, which begins on p. 9, and is taken up again on p. 239.
5. Daniel C. Dennett, "Cognitive Wheels: The Frame Problem of AI," in Christopher Hookway, ed., *Minds, Machines and Evolution* (Cambridge: Cambridge University Press, 1984), pp. 129-151. The problem was first named in John McCarthy and Patrick J. Hayes, "Some Philosophical Problems from the Standpoint of Artificial Intelligence," in B. Meltzer and D. Michie, eds., *Machine Intelligence 4* (Edinburgh: Edinburgh University Press, 1969), pp. 463-502.
6. Ludwig Wittgenstein, *Philosophical Investigations*, trans. G. E. M. Anscombe, 3rd ed. (Oxford: Basil Blackwell, 1967), §201; see also §§84-87.
7. Hubert L. Dreyfus, *What Computers Can't Do: The Limits of Artificial Intelligence*, rev. ed. (New York: Harper and Row, 1979).
8. Gilbert Ryle, *The Concept of Mind* (Harmondsworth: Penguin, 1963), chapter 2, "Knowing How and Knowing That."
9. Hubert L. Dreyfus and Stuart E. Dreyfus, *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer* (New York: Free Press, 1986), chapter 1.
10. A. M. Turing, "Computing Machinery and Intelligence," *Mind* 59 (1950), pp. 454-460, the section headed "Learning Machines."
11. Marvin Minsky, *The Society of Mind* (New York: Simon and Schuster, 1986).
12. David E. Rumelhart, James L. McClelland, and the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 2 vols. (Cambridge, Mass.: MIT Press, 1986). The learning procedure is described in vol. 1, chapter 8.
13. Terrence J. Sejnowski and Charles R. Rosenberg, *NETtalk: A Parallel Network That Learns to Read Aloud* (Baltimore: Johns Hopkins University, Department of Electrical Engineering and Computer Science, 1986), Technical Report JHU/EECS-86/01, 23 pp.
14. The model for the program is Douglas B. Lenat's AM, described in "Automated Theory Formation in Mathematics," *Proceedings of the Fifth International Joint Conference on Artificial Intelligence* (1977), pp. 833-842, and in Lenat, "AM: An Artificial Intelligence Approach to Discovery in Mathematics as Heuristic Search," Stanford University doctoral dissertation, 1976. Mine was written with AM as its guide, and is a great deal simpler.

9. Beyond the Computer: Consciousness in Matter

I said in Section 1 that the question of this paper was wider than the question about computers, and that I would use the word "machine" in its widest sense, to mean any system that is not alive, when the argument reached this point. It has reached it. Everything so far has been argued with the computer in view, because the computer is the inorganic system we know how to build and the one the objections were aimed at. But the conclusion of Section 3 was that consciousness is a matter of organization and not of material, and if that is right, then silicon is only the first alternative to carbon that we happen to have tried. The question becomes whether consciousness might be found, or built, in matter organized in ways that have nothing to do with computers at all. In this section I want to take that question as far as it will go, which is not very far, and to be honest about where it stops. I will begin with a book that takes it much further than I can, because it is the book that first put the question into my head.

Bentov's proposal

Itzhak Bentov was not a philosopher. He was an engineer and inventor who spent his working life on medical instruments, and his book *Stalking the Wild Pendulum* is an attempt by a practical man to say what consciousness is, using the tools he had, which were the physics of oscillation and a good deal of imagination.¹ I read it before I had read any of the philosophers in this paper, and I want to present it fairly, since it is easy to be unfair to a book that has no footnotes.

Bentov's starting point is that everything oscillates. Atoms vibrate; the heart beats; the body, he says, rocks very slightly on the pulse of the blood in the aorta, at a rate of a few cycles a second; the earth has its rhythms and so do the stars. He takes the pendulum as the picture of all of this, and he notices something about a pendulum that everyone knows and nobody thinks about: at the top of each swing it stops. For an instant it is at rest, and then it comes back. Bentov's speculation is that consciousness is connected with those instants of rest, that in the pause at the top of every oscillation, in every system, there is a kind of opening, and that in a meditating person, whose rhythms have been brought into step with one another, the opening widens. This is the wild pendulum of his title.

From this he goes on to a much larger claim, which is the one that concerns me here. Consciousness, he says, is not a thing that some systems have and others lack. It is a property of matter as such, present in every organized system in some degree, and the degrees form a scale. He draws the scale as a curve, with minerals near the bottom, then plants, then animals, then human beings, and the curve continuing upward past us to what

he calls the absolute.² A rock, on this view, is conscious, very slightly, in the way that a very cold thing is still warm, slightly; and the difference between the rock and the reader is a difference of quantity along one scale, not a difference of kind.

Now let me say what I think this is worth. As an argument it is worth very little, and I do not think Bentov would mind my saying so, since he does not offer it as one. There is no reason given for thinking that the pause at the top of a swing has anything to do with awareness, and the curve from rock to absolute is drawn, not derived; nothing in the book says how one would measure the quantity of consciousness in a rock, or in anything, and a claim that cannot be tested in any way is not a hypothesis but a picture. Worse, the book runs together two things this paper has been at pains to keep apart. Oscillation is organization of a kind, and a rock, with its crystal lattice, is organized. But nothing in the rock models anything. There is no map, and no mark on it, and by the account of Section 6 there is nothing it is like to be a rock because there is no place in the rock from which anything could be like anything. Bentov has mistaken being organized for being organized in the way that makes a point of view.

And yet I owe the book two things, and I want to pay them. The first is the idea that consciousness may admit of degrees. Every philosopher in this paper except Nagel treats it as a thing one has or lacks, and Nagel only hints. But if consciousness is a matter of a system's modelling itself within its model of the world, then it is plainly a matter of more and less: a more detailed model, a model that reaches further into the future and further back, a model that includes more of the system's own workings. On that account the difference between a dog and a man is a difference of degree, and so might be the difference between a man and something we have not met. Bentov's curve is drawn in the wrong place and on no evidence, but the idea that there is a curve seems to me right, and it came from him.

The second is that Bentov's picture, stripped of the pendulum, is a very old position in philosophy, and one that the arguments of this paper have not ruled out. It is called panpsychism, and it deserves a hearing on its own.

Panpsychism

The doctrine that mind is present in some degree throughout matter has been held by philosophers who cannot be dismissed. Spinoza held that every thing is at once extended and thinking, the two being one substance seen two ways.³ Leibniz held that the ultimate constituents of the world are monads, each a center of perception, and that a stone differs from a soul in the confusion of its perceptions rather than in having none.⁴ Whitehead built a whole system on the idea that the basic events of the world are "occasions of experience."⁵ And Nagel, whose bat did so much work in Section 7, has written a paper called simply

"Panpsychism," in which he argues that the doctrine follows from four premises that most materialists accept, and that the materialist who rejects it has to say which premise he gives up.⁶

Nagel's argument is worth setting out, because it puts the question in exactly the form this paper needs. The premises are these. First, that a living organism is composed of ordinary matter, the same matter that rocks are made of. Second, that the mental properties of an organism are not implied by its physical properties, in the sense that a complete physical description would not entail that there is something it is like to be it. Third, that the organism really does have mental properties; that they are not a manner of speaking. And fourth, that there are no truly emergent properties of complex systems, that whatever a whole has, it has in virtue of what its parts have and how they are arranged. From these it follows that the parts must already have something mental about them, some proto-mental property, however faint, which the arrangement brings to the level at which we notice it. Otherwise the mental would have appeared from nowhere, which the fourth premise forbids.

I have thought about which premise to give up and I am not sure. The second premise is the explanatory gap of Section 7, and I have granted it. The first and third I accept. So if I am to escape panpsychism I must deny the fourth, and say that there are truly emergent properties, that a whole can have what no part has and no arrangement of the parts' properties adds up to. Everything in Section 5 and Section 6 pushes me toward saying that. Aunt Hillary converses, and no ant does, and I said that this was not a trick but the ordinary way complex things are. But I also have to admit that Nagel's fourth premise is not about conversation. Conversing is something the ants' arrangement plainly adds up to, given what ants do; it is a program-receptive feature, in Gunderson's phrase, and emergence of that kind is no mystery. The fourth premise is about the inside, and whether the inside can emerge from an arrangement of things that have no inside is precisely the question the gap leaves open. So I cannot deny the premise with any confidence, and I record that panpsychism remains, for this paper, a live option. I do not hold it. But I cannot show it false, and the reader should know that Bentov's rock, though it has no map, is not quite as far from the argument as I made it sound.

Leibniz's mill

The classical objection to any of this, panpsychism and functionalism alike, is older than the Chinese room and I want to give it, because it shows that the room was not new and that the reply to it is not new either. Leibniz asks us to imagine a machine that thinks and feels, and to imagine it enlarged until we could walk into it as into a mill. Inside, he says, we would find only parts pushing against one another, "and never anything by which to explain

a perception."⁷ By this he means what Searle meant: that one may examine every part of a mechanism and find nothing in any part that is a thought, and that therefore thought cannot be in the mechanism.

The reply is the one given in Section 5, and I will not repeat it at length. Nothing in any brick keeps out the rain, and one may walk through a house forever, examining bricks, without finding the shelter, and the shelter is real. What Leibniz's mill shows is that a perception is not a part. It does not show that a perception is not the mill, at a level of description that the tour of the parts never reaches. And the point I want to add here is that the mill is Descartes' machine, a mechanism of parts pushing, and that the systems Section 8 ended with are not mills in that sense at all. A network of units whose links have been shaped by experience does not have a part for each thing it does, and a tour of it would find not levers and gears but a pattern of strengths spread everywhere and located nowhere. Leibniz could not walk into that. It has no inside to walk into, only an organization, and the organization is where one would have to look.

Where the wider claim stops

I have now to say what would settle the wider question, whether consciousness can be found in inorganic matter organized in ways other than a computer, and the honest answer is that I have no proposal. For computers the paper had something to work with. A computer can be given Descartes' tests, because it can be asked questions and set tasks; the criteria of Wittgenstein and Dennett apply to it, because it behaves, and its behavior can be predicted from the intentional stance or not. For matter that does not behave in that sense, a crystal, a star, a weather system, the criteria give no purchase at all. One cannot ask a star whether it is in pain. One cannot adopt the intentional stance toward it, because nothing it does is better predicted by supposing it wants something. So there is nothing to say about it, one way or the other, and the paper should not pretend otherwise.

What the paper can say is this. Consciousness, on the account given here, requires organization of a particular shape: a system that models its world, that models itself within the model, and that acts on the model and revises it. That shape is not restricted to computers, and nothing in the account restricts it to carbon. But it is a shape, and most matter is not in it. A rock is not, and neither is a star, so far as anyone can tell, and the panpsychist who says they are conscious anyway has to mean something by "conscious" that the criteria cannot reach, and I have granted that he may be right and that I cannot tell. What I can tell is that if anything is ever built out of inorganic matter that has the shape, whether it is a computer or something we have not thought of, the arguments for saying it has no point of view will be the arguments this paper has examined, and they will fail for the reasons given.

The existentialist frame returns

I want to end the section by returning to the assumption I began with, because I think it has grown into something more than an assumption over the course of the argument.

Sartre's claim was that existence precedes essence: that a human being is not a thing with a fixed nature that settles in advance what he is, but a being who becomes what he is through what he does. I took this in Section 1 as a reason to set the soul aside. I now want to notice that the whole account of consciousness in this paper has the same shape. Consciousness, on the account of Sections 3 and 6, is not a thing a system has, in the way it has a mass. It is something a system does: it models, it includes itself in the model, it acts and revises. It is a doing and not a having, and Sartre said as much of consciousness itself, in his own vocabulary, when he insisted that consciousness is not a substance but an activity, always consciousness of something, never a thing sitting inside another thing.⁸

If that is right, then the question this paper began with, whether consciousness requires living matter, is a question of the wrong shape. It asks what consciousness is made of, and a doing is not made of anything. A dance is not made of the dancers, though there is no dance without them, and one who asked what material a dance requires would be asking a question that has no answer because it has no sense. The right question is what the dancers must do. The right question about consciousness is what a system must be doing, and that question this paper has tried to answer: it must be doing the thing described in Section 6, and the account of that thing makes no mention of what does it.

I find, having got here, that I am not the first to say this, and that the first was the philosopher I spent a good part of last year on. Aristotle held that the soul is not a thing lodged in the body but the form of the body, its being organized as a living thing; and he gave an example that this paper could have used at any point. If an axe were a living thing, he says, its soul would be its being an axe, its capacity to chop; and if the eye were an animal, sight would be its soul.⁹ The soul, for Aristotle, is what the body does when it is alive, and it is no more a separate substance than sight is a thing sitting inside the eye. That is functionalism, stated twenty-three centuries before Putnam, and it was stated by a man who thought a great deal about matter and form and concluded that form was what mattered. He did not think an axe could be alive. But he thought that if one were, its life would be its chopping, and he would not have asked what the axe was made of in order to find out.

So the assumption I started with turns out to be the conclusion I end with, and I take that as a sign that it was the right assumption rather than as a proof, since one can always end where one begins by going in a circle. The reader will judge whether the path between was a circle. What I claim is only this: that a being with no fixed essence becomes what it is

through what it does, that consciousness is a way of doing, and that nothing about a way of doing says what must do it. Section 10 draws the paper together and says what I expect to happen next.

Notes to Section 9

1. Itzhak Bentov, *Stalking the Wild Pendulum: On the Mechanics of Consciousness* (New York: E. P. Dutton, 1977).
2. Ibid., chapter 5, "Quantity and Quality of Consciousness," pp. 77-92.
3. Benedict de Spinoza, *Ethics*, Part II, Proposition 7 and its Scholium, in *The Chief Works of Benedict de Spinoza*, trans. R. H. M. Elwes, vol. 2 (New York: Dover, 1951).
4. G. W. Leibniz, *Monadology* (1714), §§14-19, in *Leibniz: Philosophical Writings*, ed. G. H. R. Parkinson (London: Dent, 1973).
5. Alfred North Whitehead, *Process and Reality* (New York: Macmillan, 1929), Part I.
6. Thomas Nagel, "Panpsychism," in *Mortal Questions* (Cambridge: Cambridge University Press, 1979), pp. 181-195.
7. Leibniz, *Monadology*, §17.
8. Jean-Paul Sartre, *Being and Nothingness*, trans. Hazel E. Barnes (London: Methuen, 1957), Introduction, section III, "The Pre-reflective Cogito and the Being of the Percipere," pp. lii-liv, and section V, "The Ontological Proof," pp. lx-lxii.
9. Aristotle, *De Anima*, Book II, chapter 1, 412b10-413a3, in *Aristotle's De Anima, Books II and III*, trans. D. W. Hamlyn (Oxford: Clarendon Press, 1968). I discussed Aristotle's account of form and matter at length in two papers for Philosophy 3001 at the University of Minnesota, and I am conscious of leaning on them here.

10. Conclusion: A Hopeful Prognosis

I began by asking whether there is anything about consciousness that requires it to be made of living matter, and I said I would argue that there is not. Here is what the argument has come to.

No soul is required. I assumed this at the start, in Sartre's words, and Section 2 showed that the assumption can be argued for without him. Ryle found that Descartes' second thing was a category mistake, an invisible hurrier invented because the hurry could not be found among the parts. Wittgenstein found that the inner world, the last place the ghost could hide, is not what our words for consciousness refer to, and that the criteria for it are public: resemblance, and behavior. Between them they left the question inside materialism.

Organization suffices in principle. Of the materialist positions, the one that fits the facts is functionalism, and functionalism defines a mental state by the job it does, not by the stuff that does it. The identity theory, which ties pain to a particular chemistry, is carbon chauvinism, and it begs the question this paper asked. Hofstadter's strange loop gives functionalism what it lacked, an account of how a system comes to have a point of view: it models the world, and finds itself in the model, and the mark on the map is the self. That account makes no mention of what carries the map, and Dennett and Parfit showed that when we ask, in plain cases, where a self is, the answer is: with the pattern.

The tests are evidence and not proof, and the inside is a limit on knowing and not a fact about matter. Descartes' first test, which Turing turned into a game, gives us the same evidence about a machine that we have about each other, and no more. Gunderson was right that it is not a proof. Searle's room was a confusion of levels; the man is the hardware, and no neuron understands English. What every argument left standing was the inside, Nagel's something it is like, and Section 7 conceded it in full. But the gap between the inside and every account of it is the same gap for every theory and for every mind, my professor's included, and a gap in what we can know says nothing about what a thing must be made of. The chauvinist who uses it against machines is adding a premise, and the premise is the one already rejected.

The machines we have fail Descartes' second test, and the newest work in the field is aimed at exactly that failure. An expert system is a knife with many blades and no hand. The frame problem is the rulebook's inability to contain its own application. But a network shaped by experience is not a rulebook, and Minsky's society has no central part to run out. Whether these can be made to scale is unknown. What is known is that they are built on the principle Descartes said no machine could have, a great many parts for no action in particular, and that the second test has become a specification.

Beyond the computer, the account applies to anything organized in the right shape, and to nothing that is not. Bentov's rock has no map. Panpsychism, which says it has one anyway, could not be refuted here, and I have said so. Consciousness, on the whole account, is a doing and not a having, and a doing is not made of anything, as a dance is not made of the dancers; Sartre said this of consciousness, and Aristotle, who thought the soul was the form of the body and that an axe's soul would be its chopping, said it first.

So the door is open. I have not shown that any machine is conscious, and I do not believe that any machine now built is. I have shown that the arguments for shutting the door do not succeed, that the best theory of mind we have predicts that the door is open, and that the objections which remain are objections to our knowledge and not to the possibility. That is a hopeful conclusion, and I hold it hopefully rather than certainly, which I think is the right way to hold it.

Turing ended his paper with a prediction, and the reader is entitled to one from me, to be held against me in the year 2000. Here it is. By then a machine built on the principles of Section 8 will pass Descartes' first test with any examiner who is not a philosopher, will still fail the second, and people will have stopped asking whether it thinks and begun asking what it wants.

Acknowledgment

This essay was written over a year of twice-weekly meetings with Vernon Pratt, who set the reading, poured the coffee, and disagreed with me at every step, which is what I was there for. His own book on the history of the question was in press while this was being written, and I am conscious that whatever is clear in these pages was made clear across his table.

Bibliography

- Anderson, Alan Ross, ed. *Minds and Machines*. Englewood Cliffs, N.J.: Prentice-Hall, 1964.
- Aristotle. *De Anima, Books II and III*. Translated by D. W. Hamlyn. Oxford: Clarendon Press, 1968.
- Armstrong, D. M. *A Materialist Theory of the Mind*. London: Routledge and Kegan Paul, 1968.
- Bentov, Itzhak. *Stalking the Wild Pendulum: On the Mechanics of Consciousness*. New York: E. P. Dutton, 1977.
- Block, Ned. "Troubles with Functionalism." In *Perception and Cognition: Issues in the Foundations of Psychology*, edited by C. Wade Savage, 261-325. Minnesota Studies in the Philosophy of Science, vol. 9. Minneapolis: University of Minnesota Press, 1978.
- Boden, Margaret. *Artificial Intelligence and Natural Man*. Hassocks: Harvester Press, 1977.
- Churchland, Paul M. "Eliminative Materialism and the Propositional Attitudes." *Journal of Philosophy* 78 (1981): 67-90.
- Churchland, Paul M. *Matter and Consciousness: A Contemporary Introduction to the Philosophy of Mind*. Cambridge, Mass.: MIT Press, 1984.
- Dennett, Daniel C. *Brainstorms: Philosophical Essays on Mind and Psychology*. Hassocks: Harvester Press, 1979.
- Dennett, Daniel C. "Cognitive Wheels: The Frame Problem of AI." In *Minds, Machines and Evolution*, edited by Christopher Hookway, 129-151. Cambridge: Cambridge University Press, 1984.
- Descartes, René. *The Philosophical Works of Descartes*. Translated by Elizabeth S. Haldane and G. R. T. Ross. Vol. 1. Cambridge: Cambridge University Press, 1911.
- Dreyfus, Hubert L. *What Computers Can't Do: The Limits of Artificial Intelligence*. Rev. ed. New York: Harper and Row, 1979.
- Dreyfus, Hubert L., and Stuart E. Dreyfus. *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. New York: Free Press, 1986.
- Feigenbaum, Edward A., and Pamela McCorduck. *The Fifth Generation: Artificial Intelligence and Japan's Computer Challenge to the World*. Reading, Mass.: Addison-Wesley, 1983.
- Fodor, Jerry A. "The Mind-Body Problem." *Scientific American* 244 (January 1981): 114-123.
- Gunderson, Keith. "The Imitation Game." *Mind* 73 (1964): 234-245.
- Gunderson, Keith. *Mentality and Machines*. 2nd ed. Minneapolis: University of Minnesota Press, 1985.
- Hofstadter, Douglas R. *Gödel, Escher, Bach: An Eternal Golden Braid*. Harmondsworth: Penguin, 1980.
- Hofstadter, Douglas R., and Daniel C. Dennett, eds. *The Mind's I: Fantasies and Reflections on Self and Soul*. Harmondsworth: Penguin, 1982.
- Jackson, Frank. "Epiphenomenal Qualia." *Philosophical Quarterly* 32 (1982): 127-136.
- Jefferson, G. "The Mind of Mechanical Man." *British Medical Journal* 1 (1949): 1105-1121.

- Johnson, Jeffrey P. "Homosexuality: Is It Morally Permissible?" Paper for Philosophy 1003, University of Minnesota, 1985.
- Leibniz, G. W. *Philosophical Writings*. Edited by G. H. R. Parkinson. London: Dent, 1973.
- Lenat, Douglas B. "Automated Theory Formation in Mathematics." *Proceedings of the Fifth International Joint Conference on Artificial Intelligence* (1977): 833-842.
- Levine, Joseph. "Materialism and Qualia: The Explanatory Gap." *Pacific Philosophical Quarterly* 64 (1983): 354-361.
- Lewis, David. *Philosophical Papers*. Vol. 1. New York: Oxford University Press, 1983.
- Lucas, J. R. "Minds, Machines and Gödel." *Philosophy* 36 (1961): 112-127.
- McCarthy, John, and Patrick J. Hayes. "Some Philosophical Problems from the Standpoint of Artificial Intelligence." In *Machine Intelligence 4*, edited by B. Meltzer and D. Michie, 463-502. Edinburgh: Edinburgh University Press, 1969.
- Minsky, Marvin. *The Society of Mind*. New York: Simon and Schuster, 1986.
- Nagel, Thomas. *Mortal Questions*. Cambridge: Cambridge University Press, 1979.
- Nagel, Thomas. "What Is It Like to Be a Bat?" *Philosophical Review* 83 (1974): 435-450.
- Nemirow, Laurence. Review of *Mortal Questions*, by Thomas Nagel. *Philosophical Review* 89 (1980): 473-477.
- Parfit, Derek. *Reasons and Persons*. Oxford: Clarendon Press, 1984.
- Place, U. T. "Is Consciousness a Brain Process?" *British Journal of Psychology* 47 (1956): 44-50.
- Putnam, Hilary. *Mind, Language and Reality: Philosophical Papers*. Vol. 2. Cambridge: Cambridge University Press, 1975.
- Rumelhart, David E., James L. McClelland, and the PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. 2 vols. Cambridge, Mass.: MIT Press, 1986.
- Ryle, Gilbert. *The Concept of Mind*. Harmondsworth: Penguin, 1963.
- Sagan, Carl. *The Cosmic Connection: An Extraterrestrial Perspective*. New York: Doubleday, 1973.
- Sartre, Jean-Paul. *Being and Nothingness*. Translated by Hazel E. Barnes. London: Methuen, 1957.
- Sartre, Jean-Paul. *Existentialism and Humanism*. Translated by Philip Mairet. London: Methuen, 1948.
- Searle, John R. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (1980): 417-457.
- Searle, John R. *Minds, Brains and Science: The 1984 Reith Lectures*. London: British Broadcasting Corporation, 1984.
- Sejnowski, Terrence J., and Charles R. Rosenberg. *NETtalk: A Parallel Network That Learns to Read Aloud*. Baltimore: Johns Hopkins University, Department of Electrical Engineering and Computer Science, 1986. Technical Report JHU/EECS-86/01.
- Shortliffe, Edward H. *Computer-Based Medical Consultations: MYCIN*. New York: Elsevier, 1976.
- Sloman, Aaron. *The Computer Revolution in Philosophy*. Hassocks: Harvester Press, 1978.
- Smart, J. J. C. "Sensations and Brain Processes." *Philosophical Review* 68 (1959): 141-156.

Spinoza, Benedict de. *The Chief Works of Benedict de Spinoza*. Translated by R. H. M. Elwes. Vol. 2. New York: Dover, 1951.

Turing, A. M. "Computing Machinery and Intelligence." *Mind* 59 (1950): 433-460.

Weizenbaum, Joseph. *Computer Power and Human Reason: From Judgment to Calculation*. San Francisco: W. H. Freeman, 1976.

Weizenbaum, Joseph. "ELIZA: A Computer Program for the Study of Natural Language Communication Between Man and Machine." *Communications of the ACM* 9 (1966): 36-45.

Whitehead, Alfred North. *Process and Reality*. New York: Macmillan, 1929.

Wittgenstein, Ludwig. *Philosophical Investigations*. Translated by G. E. M. Anscombe. 3rd ed. Oxford: Basil Blackwell, 1967.