JOHN R. SEARLE

Minds, Brains, and Programs

What psychological and philosophical significance should we attach to recent efforts at computer simulations of human cognitive capacities? In answering this question, I find it useful to distinguish what I will call "strong" AI from "weak" or "cautious" AI (artificial intelligence). According to weak AI, the principal value of the computer in the study of the mind is that it gives us a very powerful tool. For example, it enables us to formulate and test hypotheses in a more rigorous and precise fashion. But according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really *is* a mind, in the sense that computers given the right programs can be literally said to *understand* and have other cognitive states. In strong AI, because the programmed computer has cognitive states, the programs are not mere tools that enable us to test psychological explanations; rather, the programs are themselves the explanations.

I have no objection to the claims of weak AI, at least as far as this article is concerned. My discussion here will be directed at the claims I have defined as those of strong AI, specifically the claim that the appropriately programmed computer literally has cognitive states and that the programs thereby explain human cognition. When I hereafter refer to AI, I have in mind the strong version, as expressed by these two claims.

I will consider the work of Roger Schank and his colleagues at Yale (Schank and Abelson 1977), because I am more familiar with it than I am with any other similar claims, and because it provides a very clear example of the sort of work I wish to examine. But nothing that follows depends upon the details of Schank's programs. The same arguments would apply to Winograd's SHRDLU (Winograd 1973), Weizenbaum's ELIZA (Weizenbaum 1965), and indeed any Turing machine simulation of human mental phenomena. See "Further Reading" for Searle's references.

Very briefly, and leaving out the various details, one can describe Schank's program as follows: The aim of the program is to simulate the human ability to understand stories. It is characteristic of human beings' story-understanding capacity that they can answer questions about the story even though the information that they give was never explicitly stated in the story. Thus, for example, suppose you are given the following story: "A man went into a restaurant and ordered a hamburger. When the hamburger arrived it was burned to a crisp, and the man stormed out of the restaurant angrily, without paying for the hamburger or leaving a tip." Now, if you are asked "Did the man eat the hamburger?" you will presumably answer, "No, he did not." Similarly, if you are given the following story: "A man went into a restaurant and ordered a hamburger; when the hamburger came he was very pleased with it; and as he left the restaurant he gave the waitress a large tip before paying his bill," and you are asked the question, "Did the man eat the hamburger?" you will presumably answer, "Yes, he ate the hamburger." Now Schank's machines can similarly answer questions about restaurants in this fashion. To do this, they have a "representation" of the sort of information that human beings have about restaurants, which enables them to answer such questions as those above, given these sorts of stories. When the machine is given the story and then asked the question, the machine will print out answers of the sort that we would expect human beings to give if told similar stories. Partisans of strong AI claim that in this question and answer sequence the machine is not only simulating a human ability but also (1) that the machine can literally be said to *understand* the story and provide the answers to questions, and (2) that what the machine and its program do *explains* the human ability to understand the story and answer questions about it.

Both claims seem to me to be totally unsupported by Schank's work, as I will attempt to show in what follows. I am not, of course, saying that Schank himself is committed to these claims.

One way to test any theory of the mind is to ask oneself what it would be like if my mind actually worked on the principles that the theory says all minds work on. Let us apply this test to the Schank program with the following *Gedankenexperiment.* Suppose that I'm locked in a room and given a large batch of Chinese writing. Suppose furthermore (as is indeed the case) that I know no Chinese, either written or spoken, and that I'm not even confident that I could recognize Chinese writing as Chinese writing distinct from, say, Japanese writing or meaningless squiggles. To me, Chinese writing is just so many meaningless squiggles. Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as well as any other native speaker of English. They enable me to correlate one set of formal symbols with another set of formal symbols, and all that "formal" means here is that I can identify the symbols entirely by their shapes. Now suppose also that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch. Unknown to me, the people who are giving me all of these symbols call the first batch a "script," they call the second batch a "story," and they call the third batch "questions." Furthermore, they call the symbols I give them back in response to the third batch "answers to the questions," and the set of rules in English that they gave me, they call the "program." Now just to complicate the story a little, imagine that these people also give me stories in English, which I understand, and they then ask me questions in English about these stories, and I give them back answers in English. Suppose also that after a while I get so good at following the instructions for manipulating the Chinese symbols and the programmers get so good at writing the programs that from the external point of view—that is, from the point of view of somebody outside the room in which I am locked—my answers to the questions are absolutely indistinguishable from those of native Chinese speakers. Nobody just looking at my answers can tell that I don't speak a word of Chinese. Let us also suppose that my answers to the English questions are, as they no doubt would be, indistinguishable from those of other native English speakers, for the simple reason that I am a native English speaker. From the external point of view—from the point of view of someone reading my "answers"—the answers to the Chinese questions and the English questions are equally good. But in the Chinese case, unlike the English case, I produce the answers by manipulating uninterpreted formal symbols. As far as the Chinese is concerned, I simply behave like a computer; I perform computational operations on formally specified elements. For the purposes of the Chinese, I am simply an instantiation of the computer program.

Now the claims made by strong AI are that the programmed computer understands the stories and that the program in some sense explains human understanding. But we are now in a position to examine these claims in light of our thought experiment.

1. As regards the first claim, it seems to me quite obvious in the example that I do not understand a word of the Chinese stories. I have inputs and outputs that are indistinguishable from those of the native Chinese speaker, and I can have any formal program you like, but I still understand nothing. For the same reasons, Schank's computer understands nothing of any stories, whether in Chinese, English, or whatever, since in the Chinese case the computer is me, and in cases where the computer is not me, the computer has nothing more than I have in the case where I understand nothing.

2. As regards the second claim, that the program explains human understanding, we can see that the computer and its program do not provide sufficient conditions of understanding since the computer and the program are functioning, and there is no understanding. But does it even provide a necessary condition or a significant contribution to understanding? One of the claims made by the supporters of strong AI is that when I understand a story in English, what I am doing is exactly the same—or perhaps more of the same—as what I was doing in manipulating the Chinese symbols. It is simply more formal symbol manipulation that distinguishes the case in English, where I do understand, from the case in Chinese, where I don't. I have not demonstrated that this claim is false, but it would certainly appear an incredible claim in the example. Such plausibility as the claim has derives from the supposition that we can construct a program that will have the same inputs and outputs as native speakers, and in addition we assume that speakers have some level of description where they are also instantiations of a program. On the basis of these two assumptions we assume that even if Schank's program isn't the whole story about understanding, it may be part of the story. Well, I suppose that is an empirical possibility, but not the slightest reason has so far been given to believe that it is true, since what is suggested—though certainly not demonstrated—by the example is that the computer program is simply irrelevant to my understanding of the story. In the Chinese case I have everything that artificial intelligence can put into me by way of a program, and I understand nothing; in the English case I

understand everything, and there is so far no reason at all to suppose that my understanding has anything to do with computer programs, that is, with computational operations on purely formally specified elements. As long as the program is defined in terms of computational operations on purely formally defined elements, what the example suggests is that these by themselves have no interesting connection with understanding. They are certainly not sufficient conditions, and not the slightest reason has been given to suppose that they are necessary conditions or even that they make a significant contribution to understanding. Notice that the force of the argument is not simply that different machines can have the same input and output while operating on different formal principles—that is not the point at all. Rather, whatever purely formal principles you put into the computer, they will not be sufficient for understanding, since a human will be able to follow the formal principles without understanding anything. No reason whatever has been offered to suppose that such principles are necessary or even contributory, since no reason has been given to suppose that when I understand English I am operating with any formal program at all.

Well, then, what is it that I have in the case of the English sentences that I do not have in the case of the Chinese sentences? The obvious answer is that I know what the former mean, while I haven't the faintest idea what the latter mean. But in what does this consist and why couldn't we give it to a machine, whatever it is? I will return to this question later, but first I want to continue with the example.

I have had the occasions to present this example to several workers in artificial intelligence, and, interestingly, they do not seem to agree on what the proper reply to it is. I get a surprising variety of replies, and in what follows I will consider the most common of these (specified along with their geographic origins).

But first I want to block some common misunderstandings about "understanding": In many of these discussions one finds a lot of fancy footwork about the word "understanding." My critics point out that there are many different degrees of understanding; that "understanding" is not a simple two-place predicate; that there are even different kinds and levels of understanding, and often the law of excluded middle doesn't even apply in a straightforward way to statements of the form "*x* understands *y*"; that in many cases it is a matter for decision and not a simple matter of fact whether *x* understands *y*; and so on. To all of these points I want to say: of course, of course. But they have nothing to do with the points at issue. There are clear cases in which "understanding" literally applies and clear cases in which it does not apply; and these two sorts of cases are all I need for this argument.* I understand stories in English; to a lesser degree I can understand stories in French; to a still lesser degree, stories in German; and in Chinese, not at all. My car and my adding machine, on the other hand, understand nothing: they are not in that line of business. We often attribute "understanding" and other cognitive predicates by metaphor and analogy to cars, adding machines, and other artifacts, but nothing is proved by such attributions. We say, "The door *knows* when to open because of its photoelectric cell," "The adding machine *knows how (understands how,* is *able)* to do addition and subtraction but not division," and "The thermostat *perceives* changes in the temperature." The reason we make these attributions is quite interesting, and it has to do with the fact that in artifacts we extend our own intentionality;† our tools are extensions of our purposes, and so we find it natural to make metaphorical attributions of intentionality to them; but I take it no philosophical ice is cut by such examples. The sense in which an automatic door "understands instructions" from its photoelectric cell is not at all the sense in which I understand English. If the sense in which Schank's programmed computers understand stories is supposed to be the metaphorical sense in which the door understands, and not the sense in which I understand English, the issue would not be worth discussing. But Newell and Simon (1963) write that the kind of cognition they claim for computers is exactly the same as for human beings. I like the straightforwardness of this claim, and it is the sort of claim I will be considering. I will argue that in the literal sense the programmed computer understands what the car and the adding machine understand, namely, exactly nothing. The computer understanding is not just (like my understanding of German) partial or incomplete; it is zero.

Now to the replies:

1. **The Systems Reply (Berkeley).** "While it is true that the individual person who is locked in the room does not understand the story, the fact is that he is merely part of a whole system, and the system does understand the story. The person has a large ledger in front of him in which are written the rules, he has a lot of scratch paper and pencils for doing calculations, he has 'data banks' of sets of Chinese symbols. Now, understanding is not being ascribed to the mere individual; rather it is being ascribed to this whole system of which he is a part."

*Also, "understanding" implies both the possession of mental (intentional) states and the truth (validity, success) of these states. For the purposes of this discussion we are concerned only with the possession of the states.

†Intentionality is by definition that feature of certain mental states by which they are directed at or about objects and states of affairs in the world. Thus, beliefs, desires, and intentions are intentional states; undirected forms of anxiety and depression are not.

My response to the systems theory is quite simple: Let the individual internalize all of these elements of the system. He memorizes the rules in the ledger and the data banks of Chinese symbols, and he does all the calculations in his head. The individual then incorporates the entire system. There isn't anything at all to the system that he does not encompass. We can even get rid of the room and suppose he works outdoors. All the same, he understands nothing of the Chinese, and a fortiori neither does the system, because there isn't anything in the system that isn't in him. If he doesn't understand, then there is no way the system could understand because the system is just a part of him.

Actually I feel somewhat embarrassed to give even this answer to the systems theory because the theory seems to me so implausible to start with. The idea is that while a person doesn't understand Chinese, somehow the *conjunction* of that person and bits of paper might understand Chinese. It is not easy for me to imagine how someone who was not in the grip of an ideology would find the idea at all plausible. Still, I think many people who are committed to the ideology of strong AI will in the end be inclined to say something very much like this; so let us pursue it a bit further. According to one version of this view, while the man in the internalized systems example doesn't understand Chinese in the sense that a native Chinese speaker does (because, for example, he doesn't know that the story refers to restaurants and hamburgers, etc.), still "the man as a formal symbol manipulation system" *really does understand Chinese.* The subsystem of the man that is the formal symbol manipulation system for Chinese should not be confused with the subsystem for English.

So there are really two subsystems in the man; one understands English, the other Chinese, and "it's just that the two systems have little to do with each other." But, I want to reply, not only do they have little to do with each other, they are not even remotely alike. The subsystem that understands English (assuming we allow ourselves to talk in this jargon of "subsystems" for a moment) knows that the stories are about restaurants and eating hamburgers, he knows that he is being asked questions about restaurants and that he is answering questions as best he can by making various inferences from the content of the story, and so on. But the Chinese system knows none of this. Whereas the English subsystem knows that "hamburgers" refers to hamburgers, the Chinese subsystem knows only that "squiggle squiggle" is followed by "squoggle squoggle." All he knows is that various formal symbols are being introduced at one end and manipulated according to rules written in English, and other symbols are going out at the other end. The whole point of the original example was to argue that such symbol manipulation by itself couldn't be sufficient for understanding Chinese in any literal sense because the man could write "squoggle squoggle" after "squiggle squiggle" without understanding anything in Chinese. And it doesn't meet that argument to postulate subsystems within the man, because the subsystems are no better off than the man was in the first place; they still don't have anything even remotely like what the English-speaking man (or subsystem) has. Indeed, in the case as described, the Chinese subsystem is simply a part of the English subsystem, a part that engages in meaningless symbol manipulation according to rules in English.

Let us ask ourselves what is supposed to motivate the systems reply in the first place; that is, what *independent* grounds are there supposed to be for saying that the agent must have a subsystem within him that literally understands stories in Chinese? As far as I can tell the only grounds are that in the example I have the same input and output as native Chinese speakers and a program that goes from one to the other. But the whole point of the examples has been to try to show that that couldn't be sufficient for understanding, in the sense in which I understand stories in English, because a person, and hence the set of systems that go to make up a person, could have the right combination of input, output, and program and still not understand anything in the relevant literal sense in which I understand English. The only motivation for saying there *must* be a subsystem in me that understands Chinese is that I have a program and I can pass the Turing test; I can fool native Chinese speakers. But precisely one of the points at issue is the adequacy of the Turing test. The example shows that there could be two "systems," both of which pass the Turing test, but only one of which understands; and it is no argument against this point to say that since they both pass the Turing test they must both understand, since this claim fails to meet the argument that the system in me that understands English has a great deal more than the system that merely processes Chinese. In short, the systems reply simply begs the question by insisting without argument that the system must understand Chinese.

Furthermore, the systems reply would appear to lead to consequences that are independently absurd. If we are to conclude that there must be cognition in me on the grounds that I have a certain sort of input and output and a program in between, then it looks like all sorts of noncognitive subsystems are going to turn out to be cognitive. For example, there is a level of description at which my stomach does information processing, and it instantiates any number of computer programs, but I take it we do not want to say that it has any understanding (cf. Pylyshyn 1980). But if we accept the systems reply, then it is hard to see how we avoid saying that stomach, heart, liver, and so on are all understanding subsystems, since there is no principled way to distinguish the motivation for saying the Chinese subsystem understands from saying that the stomach understands. It is, by the way, not an answer to this point to say that the Chinese system has information as input and output and the stomach has food and food products as input and output, since from the point of view of the agent, from my point of view, there is no information in either the food or the Chinese—the Chinese is just so many meaningless squiggles. The information in the Chinese case is solely in the eyes of the programmers and the interpreters, and there is nothing to prevent them from treating the input and output of my digestive organs as information if they so desire.

This last point bears on some independent problems in strong AI, and it is worth digressing for a moment to explain it. If strong AI is to be a branch of psychology, then it must be able to distinguish those systems that are genuinely mental from those that are not. It must be able to distinguish the principles on which the mind works from those on which nonmental systems work; otherwise it will offer us no explanations of what is specifically mental about the mental. And the mental-nonmental distinction cannot be just in the eye of the beholder but it must be intrinsic to the systems; otherwise it would be up to any beholder to treat people as nonmental and, for example, hurricanes as mental if he likes. But quite often in the AI literature the distinction is blurred in ways that would in the long run prove disastrous to the claim that AI is a cognitive inquiry. McCarthy, for example, writes. "Machines as simple as thermostats can be said to have beliefs, and having beliefs seems to be a characteristic of most machines capable of problem solving performance" (McCarthy 1979). Anyone who thinks strong AI has a chance as a theory of the mind ought to ponder the implications of that remark. We are asked to accept it as a discovery of strong AI that the hunk of metal on the wall that we use to regulate the temperature has beliefs in exactly the same sense that we, our spouses, and our children have beliefs, and furthermore that "most" of the other machines in the room—telephone, tape recorder, adding machine, electric light switch—also have beliefs in this literal sense. It is not the aim of this article to argue against McCarthy's point, so I will simply assert the following without argument. The study of the mind starts with such facts as that humans have beliefs, while thermostats, telephones, and adding machines don't. If you get a theory that denies this point you have produced a counterexample to the theory and the theory is false. One gets the impression that people in AI who write this sort of thing think they can get away with it because they don't really take it seriously, and they don't think anyone else will either. I propose, for a moment at least, to take it seriously. Think hard for one minute about what would be necessary to establish that that hunk of metal on the wall over there had real beliefs, beliefs with direction of fit, propositional content, and conditions of satisfaction; beliefs that had the possibility of being strong beliefs or weak beliefs; nervous, anxious, or secure beliefs; dogmatic, rational, or superstitious beliefs; blind faiths or hesitant cogitations; any kind of beliefs. The thermostat is not a candidate. Neither is stomach, liver, adding machine, or telephone. However, since we are taking the idea seriously, notice that its truth would be fatal to strong AI's claim to be a science of the mind. For now the mind is everywhere. What we wanted to know is what distinguishes the mind from thermostats and livers. And if McCarthy were right, strong AI wouldn't have a hope of telling us that.

2. **The Robot Reply (Yale).** "Suppose we wrote a different kind of program from Schank's program. Suppose we put a computer inside a robot, and this computer would not just take in formal symbols as input and give out formal symbols as output, but rather would actually operate the robot in such a way that the robot does something very much like perceiving, walking, moving about, hammering nails, eating, drinking—anything you like. The robot would, for example, have a television camera attached to it that enabled it to see, it would have arms and legs that enabled it to 'act,' and all of this would be controlled by its computer 'brain.' Such a robot would, unlike Schank's computer, have genuine understanding and other mental states."

The first thing to notice about the robot reply is that it tacitly concedes that cognition is not solely a matter of formal symbol manipulation, since this reply adds a set of causal relations with the outside world (cf. Fodor 1980). But the answer to the robot reply is that the addition of such "perceptual" and "motor" capacities adds nothing by way of understanding, in particular, or intentionality, in general, to Schank's original program. To see this, notice that the same thought experiment applies to the robot case. Suppose that instead of the computer inside the robot, you put me inside the room and, as in the original Chinese case, you give me more Chinese symbols with more instructions in English for matching Chinese symbols to Chinese symbols and feeding back Chinese symbols to the outside. Suppose, unknown to me, some of the Chinese symbols that come to me come from a television camera attached to the robot and other Chinese symbols that I am giving out serve to make the motors inside the robot move the robot's legs or arms. It is important to emphasize that all I am doing is manipulating formal symbols: I know none of these other facts. I am receiving "information" from the robot's "perceptual" apparatus, and I am giving out "instructions" to its motor apparatus without knowing either of these facts. I am the robot's homunculus, but unlike the traditional homunculus, I don't know what's going on. I don't understand anything except the rules for symbol manipulation. Now in this case I want to say that the robot has no intentional states at all; it is simply moving about as a result of its electrical wiring and its program. And furthermore, by instantiating the program I have no intentional states of the relevant type. All I do is follow formal instructions about manipulating formal symbols.

3. **The Brain Simulator Reply (Berkeley and M.I.T.).** "Suppose we design a program that doesn't represent information that we have about the world, such as the information in Schank's scripts, but simulates the actual sequence of neuron firings at the synapses of the brain of a native Chinese speaker when he understands stories in Chinese and gives answers to them. The machine takes in Chinese stories and questions about them as input, it simulates the formal structure of actual Chinese brains in processing these stories, and it gives out Chinese answers as outputs. We can even imagine that the machine operates, not with a single serial program, but with a whole set of programs operating in parallel, in the manner that actual human brains presumably operate when they process natural language. Now surely in such a case we would have to say that the machine understood the stories; and if we refuse to say that, wouldn't we also have to deny that native Chinese speakers understood the stories? At the level of the synapses, what would or could be different about the program of the computer and the program of the Chinese brain?"

Before countering this reply I want to digress to note that it is an odd reply for any partisan of artificial intelligence (or functionalism, etc.) to make: I thought the whole idea of strong AI is that we don't need to know how the brain works to know how the mind works. The basic hypothesis, or so I had supposed, was that there is a level of mental operations consisting of computational processes over formal elements that constitute the essence of the mental and can be realized in all sorts of different brain processes, in the same way that any computer program can be realized in different computer hardwares: On the assumptions of strong AI, the mind is to the brain as the program is to the hardware, and thus we can understand the mind without doing neurophysiology. If we had to know how the brain worked to do AI, we wouldn't bother with AI. However, even getting this close to the operation of the brain is still not sufficient to produce understanding. To see this, imagine that instead of a monolingual man in a room shuffling symbols we have the man operate an elaborate set of water pipes with valves connecting them. When the man receives the Chinese symbols, he looks up in the program, written in English, which valves he has to turn on and off. Each water connection corresponds to a synapse in the Chinese brain, and the whole system is rigged up so that after doing all the right firings, that is after turning on all the right faucets, the Chinese answers pop out at the output end of the series of pipes.

Now where is the understanding in this system? It takes Chinese as input, it simulates the formal structure of the synapses of the Chinese brain, and it gives Chinese as output. But the man certainly doesn't understand Chinese, and neither do the water pipes, and if we are tempted to adopt what I think is the absurd view that somehow the *conjunction* of man *and* water pipes understands, remember that in principle the man can internalize the formal structure of the water pipes and do all the "neuron firings" in his imagination. The problem with the brain simulator is that it is simulating the wrong things about the brain. As long as it simulates only the formal structure of the sequence of neuron firings at the synapses, it won't have simulated what matters about the brain, namely its causal properties, its ability to produce intentional states. And that the formal properties are not sufficient for the causal properties is shown by the water pipe example: we can have all the formal properties carved off from the relevant neurobiological causal properties.

4. **The Combination Reply (Berkeley and Stanford).** "While each of the previous three replies might not be completely convincing by itself as a refutation of the Chinese room counterexample, if you take all three together they are collectively much more convincing and even decisive. Imagine a robot with a brain-shaped computer lodged in its cranial cavity, imagine the computer programmed with all the synapses of a human brain, imagine the whole behavior of the robot is indistinguishable from human behavior, and now think of the whole thing as a unified system and not just as a computer with inputs and outputs. Surely in such a case we would have to ascribe intentionality to the system."

I entirely agree that in such a case we would find it rational and indeed irresistible to accept the hypothesis that the robot had intentionality, as long as we knew nothing more about it. Indeed, besides appearance and behavior, the other elements of the combination are really irrelevant. If we could build a robot whose behavior was indistinguishable over a large range from human behavior, we would attribute intentionality to it, pending some reason not to. We wouldn't need to know in advance that its computer brain was a formal analogue of the human brain.

But I really don't see that this is any help to the claims of strong AI, and here's why: According to strong AI, instantiating a formal program with the right input and output is a sufficient condition of, indeed is constitutive of, intentionality. As Newell (1979) puts it, the essence of the mental is the operation of a physical symbol system. But the attributions of intentionality that we make to the robot in this example have nothing to do with formal programs. They are simply based on the assumption that if the robot looks and behaves sufficiently like us, then we would suppose, until proven otherwise, that it must have mental states like ours that cause and are expressed by its behavior and it must have an inner mechanism capable of producing such mental states. If we knew independently how to account for its behavior without such assumptions we would not attribute intentionality to it, especially if we knew it had a formal program. And this is precisely the point of my earlier reply to objection II.

Suppose we knew that the robot's behavior was entirely accounted for by the fact that a man inside it was receiving uninterpreted formal symbols from the robot's sensory receptors and sending out uninterpreted formal symbols to its motor mechanisms, and the man was doing this symbol manipulation in accordance with a bunch of rules. Furthermore, suppose the man knows none of these facts about the robot, all he knows is which operations to perform on which meaningless symbols. In such a case we would regard the robot as an ingenious mechanical dummy. The hypothesis that the dummy has a mind would now be unwarranted and unnecessary, for there is now no longer any reason to ascribe intentionality to the robot or to the system of which it is a part (except of course for the man's intentionality in manipulating the symbols). The formal symbol manipulations go on, the input and output are correctly matched, but the only real locus of intentionality is the man, and he doesn't know any of the relevant intentional states; he doesn't, for example, *see* what comes into the robot's eyes, he doesn't *intend* to move the robot's arm, and he doesn't *understand* any of the remarks made to or by the robot. Nor, for the reasons stated earlier, does the system of which man and robot are a part.

To see this point, contrast this case with cases in which we find it completely natural to ascribe intentionality to members of certain other primate species such as apes and monkeys and to domestic animals such as dogs. The reasons we find it natural are, roughly, two: We can't make sense of the animal's behavior without the ascription of intentionality, and we can see that the beasts are made of similar stuff to ourselves—that is an eye, that a nose, this is its skin, and so on. Given the coherence of the animal's behavior and the assumption of the same causal stuff underlying it, we assume both that the animal must have mental states underlying its behavior, and that the mental states must be produced by mechanisms made out of the stuff that is like our stuff. We would certainly make similar assumptions about the robot unless we had some reason not to, but as soon as we knew that the behavior was the result of a formal program, and that the actual causal properties of the physical substance were irrelevant we would abandon the assumption of intentionality.

There are two other responses to my example that come up frequently (and so are worth discussing) but really miss the point.

5. The Other Minds Reply (Yale). "How do you know that other people understand Chinese or anything else? Only by their behavior. Now the computer can pass the behavioral tests as well as they can (in principle), so if you are going to attribute cognition to other people you must in principle also attribute it to computers."

This objection really is only worth a short reply. The problem in this discussion is not about how I know that other people have cognitive states, but rather what it is that I am attributing to them when I attribute cognitive states to them. The thrust of the argument is that it couldn't be just computational processes and their output because the computational processes and their output can exist without the cognitive state. It is no answer to this argument to feign anesthesia. In "cognitive sciences" one presupposes the reality and knowability of the mental in the same way that in physical sciences one has to presuppose the reality and knowability of physical objects.

6. The Many Mansions Reply (Berkeley). "Your whole argument presupposes that AI is only about analog and digital computers. But that just happens to be the present state of technology. Whatever these causal processes are that you say are essential for intentionality (assuming you are right), eventually we will be able to build devices that have these causal processes, and that will be artificial intelligence. So your arguments are in no way directed at the ability of artificial intelligence to produce and explain cognition."

I really have no objection to this reply save to say that it in effect trivializes the project of strong AI by redefining it as whatever artificially produces and explains cognition. The interest of the original claim made on behalf of artificial intelligence is that it was a precise, well defined thesis: mental processes are computational processes over formally defined elements. I have been concerned to challenge that thesis. If the claim is redefined so that it is no longer that thesis, my objections no longer apply because there is no longer a testable hypothesis for them to apply to.

Let us now return to the question I promised I would try to answer: Granted that in my original example I understand the English and I do not understand the Chinese, and granted therefore that the machine doesn't understand either English or Chinese, still there must be something about me that makes it the case that I understand English and a corresponding something lacking in me that makes it the case that I fail to understand Chinese. Now why couldn't we give those somethings, whatever they are, to a machine?

I see no reason in principle why we couldn't give a machine the capacity to understand English or Chinese, since in an important sense our bodies with our brains are precisely such machines. But I do see very strong arguments for saying that we could not give such a thing to a machine where the operation of the machine is defined solely in terms of computational processes over formally defined elements; that is, where the operation of the machine is defined as an instantiation of a computer program. It is not because I am the instantiation of a computer program that I am able to understand English and have other forms of intentionality (I am, I suppose, the instantiation of any number of computer programs), but as far as we know it is because I am a certain sort of organism with a certain biological (i.e., chemical and physical) structure, and this structure, under certain conditions, is causally capable of producing perception, action, understanding, learning, and other intentional phenomena. And part of the point of the present argument is that only something that had those causal powers could have that intentionality. Perhaps other physical and chemical processes could produce exactly these effects; perhaps, for example, Martians also have intentionality but their brains are made of different stuff. That is an empirical question, rather like the question whether photosynthesis can be done by something with a chemistry different from that of chlorophyll.

But the main point of the present argument is that no purely formal model will ever be sufficient by itself for intentionality because the formal properties are not by themselves constitutive of intentionality, and they have by themselves no causal powers except the power, when instantiated, to produce the next stage of the formalism when the machine is running. And any other causal properties that particular realizations of the formal model have, are irrelevant to the formal model because we can always put the same formal model in a different realization where those causal properties are obviously absent. Even if, by some miracle, Chinese speakers exactly realize Schank's program, we can put the same program in English speakers, water pipes, or computers, none of which understand Chinese, the program notwithstanding.

What matters about brain operations is not the formal shadow cast by the sequence of synapses but rather the actual properties of the sequences. All the arguments for the strong version of artificial intelligence that I have seen insist on drawing an outline around the shadows cast by cognition and then claiming that the shadows are the real thing.

By way of concluding I want to try to state some of the general philosophical points implicit in the argument. For clarity I will try to do it in a question-and-answer fashion, and I begin with that old chestnut of a question:

"Could a machine think?"

The answer is, obviously, yes. We are precisely such machines.

"Yes, but could an artifact, a man-made machine, think?"

Assuming it is possible to produce artificially a machine with a nervous system, neurons with axons and dendrites, and all the rest of it, sufficiently like ours, again the answer to the question seems to be obviously, yes. If you can exactly duplicate the causes, you could duplicate the effects. And indeed it might be possible to produce consciousness, intentionality, and all the rest of it using some other sorts of chemical principles than those that human beings use. It is, as I said, an empirical question.

"OK, but could a digital computer think?"

If by "digital computer" we mean anything at all that has a level of description where it can correctly be described as the instantiation of a computer program, then again the answer is, of course, yes, since we are the instantiations of any number of computer programs, and we can think.

"But could something think, understand, and so on *solely* in virtue of being a computer with the right sort of program? Could instantiating a program, the right program of course, by itself be a sufficient condition of understanding?"

This I think is the right question to ask, though it is usually confused with one or more of the earlier questions, and the answer to it is no.

"Why not?"

Because the formal symbol manipulations by themselves don't have any intentionality; they are quite meaningless; they aren't even *symbol* manipulations, since the symbols don't symbolize anything. In the linguistic jargon, they have only a syntax but no semantics. Such intentionality as computers appear to have is solely in the minds of those who program them and those who use them, those who send in the input and those who interpret the output.

The aim of the Chinese room example was to try to show this by showing that as soon as we put something into the system that really does have intentionality (a man), and we program him with the formal program, you can see that the formal program carries no additional intentionality. It adds nothing, for example, to a man's ability to understand Chinese.

Precisely that feature of AI that seemed so appealing—the distinction between the program and the realization—proves fatal to the claim that simulation could be duplication. The distinction between the program and its realization in the hardware seems to be parallel to the distinction between the level of mental operations and the level of brain operations. And if we could describe the level of mental operations as a formal program, then it seems we could describe what was essential about the mind without doing either introspective psychology or neurophysiology of the brain. But the equation "mind is to brain as program is to hardware" breaks down at several points, among them the following three:

First, the distinction between program and realization has the consequence that the same program could have all sorts of crazy realizations that had no form of intentionality. Weizenbaum (1976, Ch. 2), for example, shows in detail how to construct a computer using a roll of toilet paper and a pile of small stones. Similarly, the Chinese story understanding program can be programmed into a sequence of water pipes, a set of wind machines, or a monolingual English speaker, none of which thereby acquires an understanding of Chinese. Stones, toilet paper, wind, and water pipes are the wrong kind of stuff to have intentionality in the first place—only something that has the same causal powers as brains can have intentionality—and though the English speaker has the right kind of stuff for intentionality you can easily see that he doesn't get any extra intentionality by memorizing the program, since memorizing it won't teach him Chinese.

Second, the program is purely formal, but the intentional states are not in that way formal. They are defined in terms of their content, not their form. The belief that it is raining, for example, is not defined as a certain formal shape, but as a certain mental content with conditions of satisfaction, a direction of fit (see Searle 1979), and the like. Indeed the belief as such hasn't even got a formal shape in this syntactic sense, since one and the same belief can be given an indefinite number of different syntactic expressions in different linguistic systems.

Third, as I mentioned before, mental states and events are literally a product of the operation of the brain, but the program is not in that way a product of the computer.

"Well if programs are in no way constitutive of mental processes, why have so many people believed the converse? That at least needs some explanation."

I don't really know the answer to that one. The idea that computer simulations could be the real thing ought to have seemed suspicious in the first place because the computer isn't confined to simulating mental operations, by any means. No one supposes that computer simulations of a five-alarm fire will burn the neighborhood down or that a computer simulation of a rainstorm will leave us all drenched. Why on earth would anyone suppose that a computer simulation of understanding actually understood anything? It is sometimes said that it would be frightfully hard to get computers to feel pain or fall in love, but love and pain are neither harder nor easier than cognition or anything else. For simulation, all you need is the right input and output and a program in the middle that transforms the former into the latter. That is all the computer has for anything it does. To confuse simulation with duplication is the same mistake, whether it is pain, love, cognition, fires, or rainstorms.

Still, there are several reasons why AI must have seemed—and to many people perhaps still does seem—in some way to reproduce and thereby explain mental phenomena, and I believe we will not succeed in removing these illusions until we have fully exposed the reasons that give rise to them.

First, and perhaps most important, is a confusion about the notion of "information processing": many people in cognitive science believe that the human brain, with its mind, does something called "information processing," and analogously the computer with its program does information processing; but fires and rainstorms, on the other hand, don't do information processing at all. Thus, though the computer can simulate the formal features of any process whatever, it stands in a special relation to the mind and brain because when the computer is properly programmed, ideally with the same program as the brain, the information processing is identical in the two cases, and this information processing is really the essence of the mental. But the trouble with this argument is that it rests on an ambiguity in the notion of "information." In the sense in which people "process information" when they reflect, say, on problems in arithmetic or when they read and answer questions about stories, the programmed computer does not do "information processing." Rather, what it does is manipulate formal symbols. The fact that the programmer and the interpreter of the computer output use the symbols to stand for objects in the world is totally beyond the scope of the computer. The computer, to repeat, has a syntax but no semantics. Thus, if you type into the computer "2 plus 2 equals?" it will type out "4." But it has no idea that "4" means 4 or that it means anything at all. And the point is not that it lacks some second-order information about the interpretation of its first-order symbols, but rather that its first-order symbols don't have any interpretations as far as the computer is concerned. All the computer has is more symbols. The introduction of the notion of "information processing" therefore produces a dilemma: either we construe the notion of "information processing" in such a way that it implies intentionality as part of the process or we don't. If the former, then the programmed computer does not do information processing, it only manipulates formal symbols. If the latter, then, though the computer does information processing, it is only doing so in the sense in which adding machines, typewriters, stomachs, thermostats, rainstorms, and hurricanes do information processing; namely, they have a level of description at which we can describe them as taking information in at one end, transforming it, and producing information as output. But in this case it is up to outside observers to interpret the input and output as information in the ordinary sense. And no similarity is established between the computer and the brain in terms of any similarity of information processing.

Second, in much of AI there is a residual behaviorism or operationalism. Since appropriately programmed computers can have input-output patterns similar to those of human beings, we are tempted to postulate mental states in the computer similar to human mental states. But once we see that it is both conceptually and empirically possible for a system to have human capacities in some realm without having any intentionality at all, we should be able to overcome this impulse. My desk adding machine has calculating capacities, but no intentionality, and in this paper I have tried to show that a system could have input and output capabilities that duplicated those of a native Chinese speaker and still not understand Chinese, regardless of how it was programmed. The Turing test is typical of the tradition in being unashamedly behavioristic and operationalistic, and I believe that if AI workers totally repudiated behaviorism and operationalism much of the confusion between simulation and duplication would be eliminated.

Third, this residual operationalism is joined to a residual form of dualism; indeed strong AI only makes sense given the dualistic assumption that, where the mind is concerned, the brain doesn't matter. In strong AI (and in functionalism, as well) what matters are programs, and programs are independent of their realization in machines; indeed, as far as AI is concerned, the same program could be realized by an electronic machine, a Cartesian mental substance, or a Hegelian world spirit. The single most surprising discovery that I have made in discussing these issues is that many AI workers are quite shocked by my idea that actual human mental phenomena might be dependent on actual physical-chemical properties of actual human brains. But if you think about it a minute you can see that I should not have been surprised; for unless you accept some form of dualism, the strong AI project hasn't got a chance. The project is to reproduce and explain the mental by designing programs, but unless the mind is not only conceptually but empirically independent of the brain you couldn't carry out the project, for the program is completely independent of any realization. Unless you believe that the mind is separable from the brain both conceptually and empirically—dualism in a strong form—you cannot hope to reproduce the mental by writing and running programs since programs must be independent of brains or any other particular forms of instantiation. If mental operations consist in computational operations on formal symbols, then it follows that they have no interesting connection with the brain; the only connection would be that the brain just happens to be one of the indefinitely many types of machines capable of instantiating the program. This form of dualism is not the traditional Cartesian variety that claims there are two sorts of *substances,* but it is Cartesian in the sense that it insists that what is specifically mental about the mind has no intrinsic connection with the actual properties of the brain. This underlying dualism is masked from us by the fact that AI literature contains frequent fulminations against "dualism"; what the authors seem to be unaware of is that their position presupposes a strong version of dualism.

"Could a machine think?" My own view is that *only* a machine could think, and indeed only very special kinds of machines, namely brains and machines that had the same causal powers as brains. And that is the main reason strong AI has had little to tell us about thinking, since it has nothing to tell us about machines. By its own definition, it is about programs, and programs are not machines. Whatever else intentionality is, it is a biological phenomenon, and it is as likely to be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomena. No one would suppose that we could produce milk and sugar by running a computer simulation of the formal sequences in lactation and photosynthesis, but where the mind is concerned many people are willing to believe in such a miracle because of a deep and abiding dualism: the mind they suppose is a matter of formal processes and is independent of quite specific material causes in the way that milk and sugar are not.

In defense of this dualism the hope is often expressed that the brain is a digital computer (early computers, by the way, were often called "electronic brains"). But that is no help. Of course the brain is a digital computer. Since everything is a digital computer, brains are too. The point is that the brain's causal capacity to produce intentionality cannot consist in its instantiating a computer program, since for any program you like it is possible for something to instantiate that program and still not have any mental states. Whatever it is that the brain does to produce intentionality, it cannot consist in instantiating a program since no program, by itself, is sufficient for intentionality.*

Reflections

This article originally appeared together with twenty-eight responses from assorted people. Many of the responses contained excellent commentary, but reprinting them would have overloaded this book, and in any case some were a little too technical. One of the nice things about Searle's article is that it is pretty much understandable by someone without special training in AI, neurology, philosophy, or other disciplines that have a bearing on it.

Our position is quite opposed to Searle's, but we find in Searle an eloquent opponent. Rather than attempt to give a thorough rebuttal to his points, we will concentrate on a few of the issues he raises, leaving our answers to his other points implicit, in the rest of this book.

Searle's paper is based on his ingenious "Chinese room thought experiment," in which the reader is urged to identify with a human being executing by hand the sequence of steps that a very clever AI program would allegedly go through as it read stories in Chinese and answered questions about them in Chinese in a manner sufficiently human-seeming as to be able to pass the Turing test. We think Searle has committed a serious and fundamental misrepresentation by giving the impression that it makes any sense to think that a human being could do this. By buying this image, the reader is unwittingly sucked into an impossibly unrealistic concept of the relation between intelligence and symbol manipulation.

The illusion that Searle hopes to induce in readers (naturally he doesn't think of it as an illusion!) depends on his managing to make readers overlook a tremendous difference in complexity between two systems at different conceptual levels. Once he has done that, the rest is a piece of cake. At the outset, the reader is invited to identify with Searle

*I am indebted to a rather large number of people for discussion of these matters and for their patient attempts to overcome my ignorance of artificial intelligence. I would especially like to thank Ned Block, Hubert Dreyfus, John Haugeland, Roger Schank, Robert Wilensky, and Terry Winograd.

as he hand-simulates an existing AI program that can, in a limited way, answer questions of a limited sort, in a few limited domains. Now, for a person to hand-simulate this, or any currently existing AI program—that is, to step through it at the level of detail that the computer does—would involve days, if not weeks or months, of arduous, horrendous boredom. But instead of pointing this out, Searle—as deft at distracting the reader's attention as a practiced magician—switches the reader's image to a hypothetical program that passes the Turing test! He has jumped up many levels of competency without so much as a passing mention. The reader is again invited to put himself or herself in the shoes of the person carrying out the step-by-step simulation, and to "feel the lack of understanding" of Chinese. This is the crux of Searle's argument.

Our response to this (and, as we shall show later, Searle's response as well, in a way) is basically the "Systems Reply": that it is a mistake to try to impute the understanding to the (incidentally) animate simulator; rather it belongs to the system as a whole, which includes what Searle casually characterizes as "a few slips of paper." This offhand comment, we feel, reveals how Searle's image has blinded him to the realities of the situation. A thinking computer is as repugnant to John Searle as non-Euclidean geometry was to its unwitting discoverer, Gerolamo Saccheri, who thoroughly disowned his own creation. The time—the late 1700s—was not quite ripe for people to accept the conceptual expansion caused by alternate geometries. About fifty years later, however, non-Euclidean geometry was rediscovered and slowly accepted.

Perhaps the same will happen with "artificial intentionality"—if it is ever created. If there ever came to be a program that could pass the Turing test, it seems that Searle, instead of marveling at the power and depth of that program, would just keep on insisting that it lacked some marvelous "causal powers of the brain" (whatever they are). To point out the vacuity of that notion, Zenon Pylyshyn, in his reply to Searle, wondered if the following passage, quite reminiscent of Zuboff's "Story of a Brain" (selection 12), would accurately characterize Searle's viewpoint:

> If more and more of the cells in your brain were to be replaced by integrated circuit chips, programmed in such a way as to keep the input-output *function* of each unit identical to that of the unit being replaced, you would in all likelihood just keep right on speaking exactly as you are doing now except that you would eventually stop *meaning* anything by it. What we outside observers might take to be words would become for you just certain noises that circuits caused you to make.

The weakness of Searle's position is that he offers no clear way to tell when genuine meaning—or indeed the genuine "you"—has vanished from this system. He merely insists that some systems have intentionality by virtue of their "causal powers" and that some don't. He vacillates about what those powers are due to. Sometimes it seems that the brain is composed of "the right stuff," but other times it seems to be something else. It is whatever seems convenient at the moment—now it is the slippery essence that distinguishes "form" from "content," now another essence that separates syntax from semantics, and so on.

To the Systems-Reply advocates, Searle offers the thought that the human being in the room (whom we shall from now on refer to as "Searle's demon") should simply memorize, or incorporate, all the material on the "few slips of paper." As if a human being could, by any conceivable stretch of the imagination, do this. The program on those "few slips of paper" embodies the entire mind and character of something as complex in its ability to respond to written material as a human being is, by virtue of being able to pass the Turing test. Could any human being simply "swallow up" the entire description of another human being's mind? We find it hard enough to memorize a written paragraph; but Searle envisions the demon as having absorbed what in all likelihood would amount to millions, if not billions, of pages densely covered with abstract symbols—and moreover having all of this information available, whenever needed, with no retrieval problems. This unlikely aspect of the scenario is all lightly described, and it is not part of Searle's key argument to convince the reader that it makes sense. In fact, quite the contrary—a key part of his argument is in glossing over these questions of orders of magnitude, for otherwise a skeptical reader will realize that nearly all of the understanding must lie in the billions of symbols on paper, and practically none of it in the demon. The fact that the demon is animate is an irrelevant—indeed, misleading—side issue that Searle has mistaken for a very significant fact.

We can back up this argument by exhibiting Searle's own espousal of the Systems Reply. To do so, we should first like to place Searle's thought experiment in a broader context. In particular, we would like to show how Searle's setup is just one of a large family of related thought experiments, several of which are the topics of other selections in this book. Each member of this family of thought experiments is defined by a particular choice of "knob settings" on a thought-experiment generator. Its purpose is to create—in your mind's eye—various sorts of imaginary simulations of human mental activity. Each different thought experiment is an "intuition pump" (Dennett's term) that magnifies one facet or other of the issue, tending to push the reader toward certain conclusions. We see approximately five knobs of interest, although it is possible that someone else could come up with more.

Knob 1. This knob controls the physical "stuff" out of which the simulation will be constructed. Its settings include: neurons and chemicals; water pipes and water; bits of paper and symbols on them; toilet paper and stones; data structures and procedures; and so on.

Knob 2. This knob controls the level of accuracy with which the simulation attempts to mimic the human brain. It can be set at an arbitrarily fine level of detail (particles inside atoms), at a coarser level such as that of cells and synapses, or even at the level that AI researchers and cognitive psychologists deal with: that of concepts and ideas, representations and processes.

Knob 3. This knob controls the physical size of the simulation. Our assumption is that microminiaturization would allow us to make a teeny-weeny network of water pipes or solid-state chips that would fit inside a thimble, and conversely that any chemical process could be blown up to the macroscopic scale.

Knob 4. This critical knob controls the size and nature of the demon who carries out the simulation. If it is a normal-sized human being, we shall call it a "Searle's demon." If it is a tiny elflike creature that can sit inside neurons or on particles, then we shall call it a "Haugeland's demon," after John Haugeland, whose response to Searle featured this notion. The settings of this knob also determine whether the demon is animate or inanimate.

Knob 5. This knob controls the speed at which the demon works. It can be set to make the demon work blindingly fast (millions of operations per microsecond) or agonizingly slowly (maybe one operation every few seconds).

Now, by playing with various knob settings, we can come up with various thought experiments. One choice yields the situation described in selection 26, "A Conversation with Einstein's Brain." Another choice yields Searle's Chinese room experiment. In particular, that involves the following knob settings:

Knob 1: paper and symbols
Knob 2: concepts and ideas
Knob 3: room size
Knob 4: human-sized demon
Knob 5: slow setting (one operation every few seconds)

Note that in principle Searle is not opposed to assuming that a simulation with these parameters could pass the Turing test. His dispute is only with what that would imply.

There is one final parameter that is not a knob but a point of view from which to look at the experiment. Let us add a little color to this drab experiment and say that the simulated Chinese speaker involved is a woman and that the demons (if animate) are always male. Now we have a choice between the demon's-eye view and the system's-eye view. Remember that by hypothesis, both the demon and the simulated woman are equally capable of articulating their views on whether or not they are understanding, and on what they are experiencing. Searle is insistent, nonetheless, that we see this experiment only from the point of view of the demon. He insists that no matter what the simulated woman claims (in Chinese, of course) about her understanding, we should disregard her claims, and pay attention to the demon inside, who is carrying out the symbol manipulation. Searle's claim amounts to the notion that actually there is only one point of view, not two. If one accepts the way Searle describes the whole experiment, this claim has great intuitive appeal, since the demon is about our size, speaks our language, and works at about our speed—and it is very hard to identify with a "woman" whose answers come at the rate of one per century (with luck)—and in "meaningless squiggles and squoggles," to boot.

But if we change some of the knob settings, we can also alter the ease with which we change point of view. In particular, Haugeland's variation involves switching various knobs as follows:

Knob 1: neurons and chemicals
Knob 2: neural-firing level
Knob 3: brain size
Knob 4: eensy-weensy demon
Knob 5: dazzlingly fast demon

What Haugeland wants us to envision is this: A real woman's brain is, unfortunately, defective. It no longer is able to send neurotransmitters from one neuron to another. Luckily, however, this brain is inhabited by an incredibly tiny and incredibly speedy Haugeland's demon, who intervenes every single time any neuron would have been about to release neurotransmitters into a neighboring neuron. This demon "tickles" the appropriate synapse of the next neuron in a way that is functionally indistinguishable, to that neuron, from the arrival of genuine neurotransmitters. And the H-demon is so swift that he can jump around from synapse to synapse in trillionths of a second, never falling behind schedule. In this way the operation of the woman's brain proceeds exactly as it would have, if she were healthy. Now, Haugeland asks Searle, does the woman still think—that is, does she possess intentionality—or, to recall the words of Professor Jefferson as cited by Turing, does she merely "artificially signal"?

You might expect Searle to urge us to listen to and identify with the demon, and to eschew the Systems Reply, which would be, of course, to listen to and identify with the woman. But in his response to Haugeland, Searle surprises us—he chooses to listen to *her* this time and to ignore the demon who is cursing us from his tiny vantage point, yelling up to us, "Fools! Don't listen to her! She's merely a puppet whose every action is caused by my tickling, and by the program embedded in these many neurons that I zip around among." But Searle does not heed the H-demon's warning cries. He says, "Her neurons still have the right causal powers; they just need some help from the demon."

We can construct a mapping between Searle's original setup and this modified setup. To the "few slips of paper" now correspond all the synapses in the woman's brain. To the AI program written on these "few slips of paper" corresponds the entire configuration of the woman's brain; this amounts to a gigantic prescription telling the demon when and how to know which synapses to tickle. To the act of writing "meaningless squiggles and squoggles of Chinese" on paper corresponds the act of tickling her synapses. Suppose we take the setup as is, except that we'll vary the size and speed knobs. We'll blow the woman's brain up to the size of the Earth, so that the demon becomes an "us-sized" S-demon instead of a tiny H-demon. And let's also have the S-demon act at speeds reasonable for humans, instead of zipping thousands of miles throughout this bulbous brain in mere microseconds. Now which level does Searle wish us to identify with? We won't speculate, but it seems to us that if the Systems Reply was compelling in the previous case, it should still be so in this case.

It must be admitted that Searle's thought experiment vividly raises the question of what understanding a language really is. We would like to digress for a moment on that topic. Consider the question: "What kind of ability to manipulate the written or spoken symbols of a language amounts to a *true understanding* of that language?" Parrots who parrot English do not understand English. The recorded voice of a woman announcing the exact time of day on the telephone time service is not the mouthpiece of a system that understands English. There is no mentality behind that voice—it has been skimmed off of its mental substrate, yet retains a human-seeming quality. Perhaps a child would wonder how anyone could have so boring a job, and could do it so reliably. This would amuse us. It would be another matter, of course, if her voice were being driven by a flexible AI program that could pass the Turing test!

Imagine you are teaching a class in China. Further, imagine that you are aware of formulating all your thoughts in English and then of applying last-minute transformation rules (in reality, they would be last-splitsecond rules) that convert the English thoughts into instructions for moving your mouth and vocal cords in strange, "meaningless" ways—and yet, all your pupils sit there and seem quite satisfied with your performance. When they raise their hands, they utter exotic sounds that, although they are completely meaningless to you, you are equipped to deal with, as you quickly apply some inverse rules and recover the English meanings underlying them. . . . Would you feel you were actually speaking Chinese? Would you feel you had gained some insight into the Chinese mentality? Or—can you actually imagine this situation? Is it realistic? Could anyone actually speak a foreign language well using this method?

The standard line is "You must learn to *think in Chinese.*" But in what does this consist? Anyone who has experienced it will recognize this description: The sounds of the second language pretty soon become "unheard"—you hear right through them, rather than hearing them, as you see right through a window, rather than seeing the window. Of course, you can make yourself hear a familiar language as pure uninterpreted sound if you try very hard, just as you can look at a windowpane if you want; but you can't have your cake and eat it too—you can't hear the sounds both *with* and *without* their meanings. And so most of the time people hear mainly meaning. For those people who learn a language because of enchantment with its sounds, this is a bit disappointing—and yet mastery of those sounds, even if one no longer hears them naively, is a beautiful, exhilarating experience. (It would be an interesting thing to try to apply this same kind of analysis to the hearing of music, where the distinction between hearing bare sounds and hearing their "meanings" is far less well understood, yet seems very real.)

Learning a second language involves transcending one's own native language. It involves mixing the new language right in with the medium in which thought takes place. Thoughts must be able to germinate as easily (or nearly as easily) in the new language as in one's native language. The way in which a new language's habits seep down level by level and finally get absorbed into neurons is a giant mystery still. But one thing for certain is that mastery of a language does not consist in getting your "English subsystem" to execute for you a program of rules that enable you to deal with a language as a set of meaningless sounds and marks. Somehow, the new language must fuse with your internal representational system—your repertoire of concepts, images, and so on—in the same intimate way as English is fused with it. To think precisely about this, one must develop a very clear notion of the concept of *levels of implementation,* a computer-science concept of great power.

Computer scientists are used to the idea that one system can "emulate" another system. In fact, it follows from a theorem proven in 1936 by Alan Turing that any general-purpose digital computer can take on the guise of any other general-purpose digital computer, and the only difference to the outside world will be one of speed. The verb "emulate" is reserved for simulations, by a computer, of another computer, while "simulate" refers to the modeling of other phenomena, such as hurricanes, population curves, national elections, or even computer users.

A major difference is that simulation is almost always approximate, depending on the nature of the model of the phenomenon in question, whereas emulation is in a deep sense exact. So exact is it that when, say, a Sigma-5 computer emulates a computer with different architecture—say a DEC PDP-10—the users of the machine will be unaware that they are not dealing with a genuine DEC. This embedding of one architecture in another gives rise to so-called "virtual machines"—in this case, a virtual DEC-10. Underneath every virtual machine there is always some other machine. It may be a machine of the same type, it may even be another virtual machine. In his book *Structured Computer Organization,* Andrew Tanenbaum uses this notion of virtual machines to explain how large computer systems can be seen as a stack of virtual machines implemented one on top of the other—the bottommost one being, of course, a *real* machine! But in any case, the levels are sealed off from each other in a watertight way, just as Searle's demon was prevented from talking to the Chinese speaker he was part of. (It is intriguing to imagine what kind of conversation would take place—assuming that there were an interpreter present, since Searle's demon knows no Chinese!)

Now in theory, it is possible to have any two such levels communicate with each other, but this has traditionally been considered bad style; level-mingling is forbidden. Nonetheless, it is probable that this forbidden fruit—this blurring of two implementational levels—is exactly what goes on when a human "system" learns a second language. The second language does not run on top of the first one as a kind of software parasite, but rather becomes equally fundamentally implanted in the hardware (or nearly so). Somehow, absorption of a second language involves bringing about deep changes in one's underlying "machine"—a vast and coherent set of changes in the ways that neurons fire, so sweeping a set of changes that it creates new ways for the higher-level entities—the symbols—to trigger one another.

To parallel this in a computer system, a higher-level program would have to have some way of creating changes inside the "demon" that is carrying its program out. This is utterly foreign to the present style in computer science of implementing one level above another in a strictly vertical, sealed-off fashion. The ability of a higher level to loop back and affect lower levels—its own underpinnings—is a kind of magic trick which we feel is very close to the core of consciousness. It will perhaps one day prove to be a key element in the push toward ever-greater flexibility in computer design, and of course in the approach toward artificial intelligence. In particular, a satisfactory answer to the question of what "understanding" really means will undoubtedly require a much sharper delineation of the ways in which different levels in a symbol-manipulating system can depend on and affect one another. All in all, these concepts have proven elusive, and a clear understanding of them is probably a good ways off yet.

In this rather confusing discussion of many levels, you may have started to wonder what in the world "level" really means. It is a most difficult question. As long as levels are sealed off from each other, like Searle's demon and the Chinese-speaking woman, it is fairly clear. When they begin to blur, beware! Searle may admit that there are two *levels* in his thought experiment, but he is reluctant to admit that there are two occupied *points of view*—two genuine beings that feel and "have experience." He is worried that once we admit that *some* computational systems might have experiences, that would be a Pandora's box and all of a sudden "mind would be everywhere"—in the churning of stomachs, livers, automobile engines, and so on.

Searle seems to believe that any system whatsoever can be *ascribed* beliefs and feelings and so on, if one looks hard enough for a way to describe the system as an instantiation of an AI program. Obviously, that would be a disturbing notion, leading the way to panpsychism. Indeed, Searle believes that the AI people have unwittingly committed themselves to a panpsychic vision of the world.

Searle's escape from his self-made trap is to maintain that all those "beliefs" and "feelings" that you will uncover in inanimate objects and so forth when you begin seeing mind everywhere are not genuine but "pseudo." They lack intentionality! They lack the causal powers of the brain! (Of course, Searle would caution others to beware of confusing *these* notions with the naïvely dualistic notion of "soul.")

Our escape is to deny that the trap exists at all. It is incorrect to see minds everywhere. We say: minds do not lurk in car engines or livers any more than brains lurk in car engines and livers.

It is worthwhile expanding on this a little. If you can see all the complexity of thought processes in a churning stomach, then what's to prevent you from reading the pattern of bubbles in a carbonated beverage as coding for the Chopin piano concerto in E minor? And don't the holes in pieces of Swiss cheese code for the entire history of the United States? Sure they do—in Chinese as well as in English. After all, all things are written everywhere! Bach's Brandenburg concerto no. 2 is coded for in the structure of Hamlet—and Hamlet was of course readable (if you'd only known the code) from the structure of the last piece of birthday cake you gobbled down.

The problem is, in all these cases, that of specifying the code without knowing in advance what you want to read. For otherwise, you could pull a description of anyone's mental activity out of a baseball game or a blade of grass by an arbitrarily constructed *a posteriori* code. But this is not science.

Minds come in different grades of sophistication, surely, but minds worth calling minds exist only where sophisticated representational systems exist, and no describable mapping that remains constant in time will reveal a self-updating representational system in a car engine or a liver. Perhaps one could read mentality into a rumbling car engine in somewhat the way that people read extra meanings into the structures of the Great Pyramids or Stonehenge, the music of Bach, Shakespeare's plays, and so on—namely, by fabricating far-fetched numerological mapping schemes that can be molded and flexed whenever needed to fit the desires of the interpreter. But we doubt that that is what Searle intends (we do grant that he intends).

Minds exist in brains and may come to exist in programmed machines. If and when such machines come about, their causal powers will derive not from the substances they are made of, but from their design and the programs that run in them. And the way we will know they have those causal powers is by talking to them and listening carefully to what they have to say.

D.R.H.