From Scientific American. (c)January 1990 by Scientific American, Inc. Permission to reprint granted by the publisher.

Is the Brain's Mind a Computer Program?

No. A program merely manipulates symbols, whereas a brain attaches meaning to them

by John R. Searle

Can a machine think? Can a machine have conscious thoughts in exactly the same sense that you and I have? If by "machine" one means a physical system capable of performing certain functions (and what else can one mean?), then humans are machines of a special biological kind, and humans can think, and so of course machines can think. And, for all we know, it might be possible to produce a thinking machine out of different materials altogether—say, out of silicon chips or vacuum tubes. Maybe it will turn out to be impossible, but we certainly do not know that yet.

In recent decades, however, the question of whether a machine can think has been given a different interpretation entirely. The question that has been posed in its place is, Could a machine think just by virtue of implementing a computer program? Is the program by itself constitutive of thinking? This is a completely different question because it is not about the physical, causal properties of actual or possible physical systems but rather about the abstract, computational properties of formal computer programs that can be implemented in any sort of substance at all, provided only that the substance is able to carry the program.

A fair number of researchers in artificial intelligence (AI) believe the answer to the second question is yes; that is, they believe that by designing the right programs with the right inputs and outputs, they are literally creating minds. They believe furthermore that they have a scientific test for determining success or failure: the Turing test devised by Alan M. Turing, the founding father of artificial intelligence. The Turing test, as currently understood, is simply this: if a computer can perform in such a way that an expert cannot distinguish its performance from that of a human who has a certain cognitive ability—say, the ability to do addition or to understand Chinese—then the computer also has that ability. So the goal is to design programs that will simulate human cognition in such a way as to pass the Turing test. What is more, such a program would not merely be a model of the mind; it would literally be a mind, in the same sense that a human mind is a mind.

By no means does every worker in artificial intelligence accept so extreme a view. A more cautious approach is to think of computer models as being useful in studying the mind in the same way that they are useful in studying the weather, economics or molecular biology. To distinguish these two approaches, I call the first strong AI and the second weak AI. It is important to see just how bold an approach strong AI is. Strong AI claims that thinking is merely the manipulation of formal symbols, and that is exactly what the computer does: manipulate formal symbols. This view is often summarized by saying, "The mind is to the brain as the program is to the hardware."

Strong AI is unusual among theories of the mind in at least two respects: it can be stated clearly, and it admits of a simple and decisive refutation. The refutation is one that any person can try for himself or herself. Here is how it goes. Consider a language you don't understand. In my case, I do not understand Chinese. To me Chinese writing looks like so many meaningless squiggles. Now suppose I am placed in a room containing baskets full of Chinese symbols. Suppose also that I am given a rule book in English for matching Chinese symbols with other Chinese symbols. The rules identify the symbols entirely by their shapes and do not require that I understand any of them. The rules might say such things as, "Take a squiggle-squiggle sign from basket number one and put it next to a squoggle-squoggle sign from basket number two."

Imagine that people outside the room who understand Chinese hand in small bunches of symbols and that in response I manipulate the symbols according to the rule book and hand back more small bunches of symbols. Now, the rule book is the "computer program." The people who wrote it are "programmers," and I am the "computer." The baskets full of symbols are the "data base," the small bunches that are handed in to me are "questions" and the bunches I then hand out are "answers."

Now suppose that the rule book is written in such a way that my "answers" to the "questions" are indistinguishable from those of a native Chinese speaker. For example, the people outside might hand me some symbols that unknown to me mean, "What's your favorite color?" and I might after going through the rules give back symbols that, also unknown to me, mean, "My favorite is blue, but I also like green a lot." I satisfy the Turing test for understanding Chinese. All the same, I am totally ignorant of Chinese. And there is no way I could come to understand Chinese in the system as described, since there is no way that I can learn the meanings of any of the symbols. Like a computer, I manipulate symbols, but I attach no meaning to the symbols.

The point of the thought experiment is this: if I do not understand Chinese solely on the basis of running a computer program for understanding Chinese, then neither does any other digital computer solely on that basis. Digital computers merely manipulate formal symbols according to rules in the program.

What goes for Chinese goes for other forms of cognition as well. Just manipulating the symbols is not by itself enough to guarantee cognition, perception, understanding, thinking and so forth. And since computers, qua computers, are symbol-manipulating devices, merely running the computer program is not enough to guarantee cognition.

JOHN R. SEARLE is professor of philosophy at the University of California, Berkeley. He received his B.A., M.A. and D.Phil. from the University of Oxford, where he was a Rhodes scholar. He wishes to thank Stuart Dreyfus, Stevan Harnad, Elizabeth Lloyd and Irvin Rock for their comments and suggestions.

This simple argument is decisive against the claims of strong AI. The first premise of the argument simply states the formal character of a computer program. Programs are defined in terms of symbol manipulations, and the symbols are purely formal, or "syntactic." The formal character of the program, by the way, is what makes computers so powerful. The same program can be run on an indefinite variety of hardwares, and one hardware system can run an indefinite range of computer programs. Let me abbreviate this "axiom" as

Axiom 1. *Computer programs are formal (syntactic).*

This point is so crucial that it is worth explaining in more detail. A digital computer processes information by first encoding it in the symbolism that the computer uses and then manipulating the symbols through a set of precisely stated rules. These rules constitute the program. For example, in Turing's early theory of computers, the symbols were simply 0's and 1's, and the rules of the program said such things as, "Print a 0 on the tape, move one square to the left and erase a 1." The astonishing thing about computers is that any information that can be stated in a language can be encoded in such a system, and any information-processing task that can be solved by explicit rules can be programmed.

Two further points are important. First, symbols and programs are purely abstract notions: they have no essential physical properties to define them and can be implemented in any physical medium whatsoever. The 0's and 1's, qua symbols, have no essential physical properties and a fortiori have no physical, causal properties. I emphasize this point because it is tempting to identify computers with some specific technology—say, silicon chips—and to think that the issues are about the physics of silicon chips or to think that syntax identifies some physical phenomenon that might have as yet unknown causal powers, in the way that actual physical phenomena such as electromagnetic radiation or hydrogen atoms have physical, causal properties. The second point is that symbols are manipulated without reference to any meanings. The symbols of the program can stand for anything the programmer or user wants. In this sense the program has syntax but no semantics.

The next axiom is just a reminder of the obvious fact that thoughts, perceptions, understandings and so forth have a mental content. By virtue of their content they can be about objects and states of affairs in the world. If the content involves language, there will be syntax in addition to semantics, but linguistic understanding requires at least a semantic framework. If, for example, I am thinking about the last presidential election, certain words will go through my mind, but the words are about the election only because I attach specific meanings to these words, in accordance with my knowledge of English. In this respect they are unlike Chinese symbols for me. Let me abbreviate this axiom as

Axiom 2. *Human minds have mental contents (semantics).*

Now let me add the point that the Chinese room demonstrated. Having the symbols by themselves—just having the syntax—is not sufficient for having the semantics. Merely manipulating symbols is not enough to guarantee knowledge of what they mean. I shall abbreviate this as

Axiom 3. *Syntax by itself is neither constitutive of nor sufficient for semantics.*

At one level this principle is true by definition. One might, of course, define the terms syntax and semantics differently. The point is that there is a distinction between formal elements, which have no intrinsic meaning or content, and those phenomena that have intrinsic content. From these premises it follows that

Conclusion 1. *Programs are neither constitutive of nor sufficient for minds.*

And that is just another way of saying that strong AI is false.

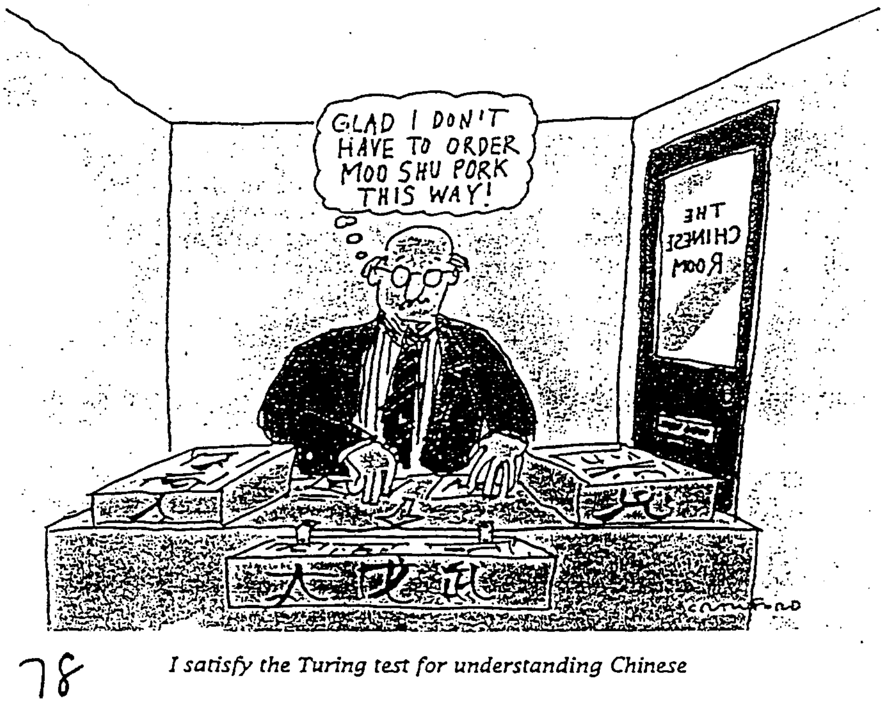
It is important to see what is proved and not proved by this argument.

First, I have not tried to prove that "a computer cannot think." Since anything that can be simulated computationally can be described as a computer, and since our brains can at some levels be simulated, it follows trivially that our brains are computers and they can certainly think. But from the fact that a system can be simulated by symbol manipulation and the fact that it is thinking, it does not follow that thinking is equivalent to formal symbol manipulation.

Second, I have not tried to show that only biologically based systems like our brains can think. Right now those are the only systems we know for a fact can think, but we might find other systems in the universe that can produce conscious thoughts, and we might even come to be able to create thinking systems artificially. I regard this issue as up for grabs.

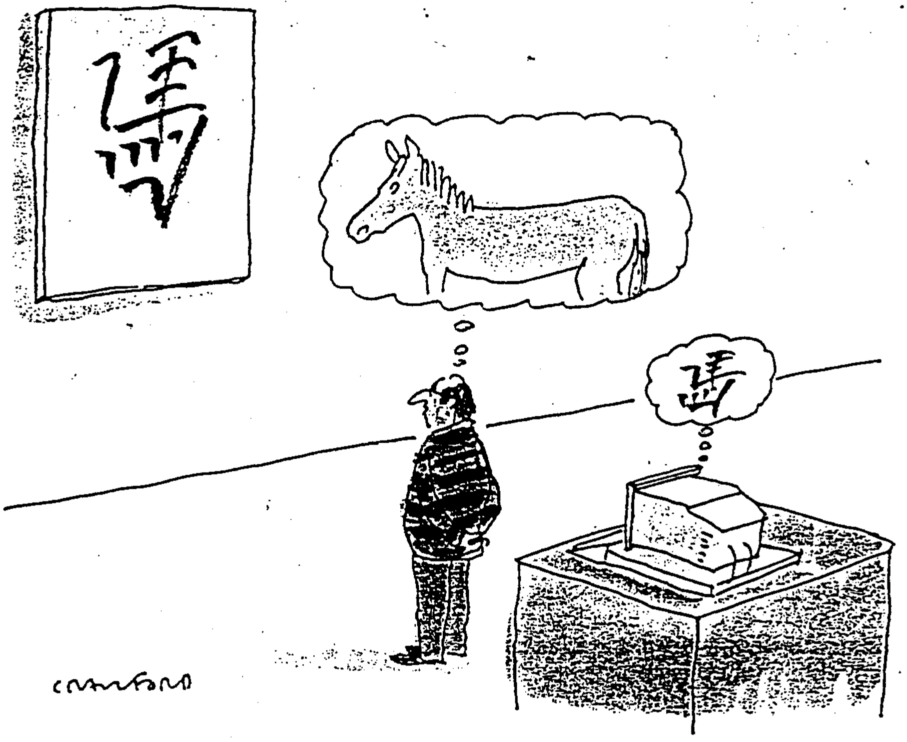
Third, strong AI's thesis is not that, for all we know, computers with the right programs might be thinking, that they might have some as yet undetected psychological properties; rather it is that they must be thinking because that is all there is to thinking.

Fourth, I have tried to refute strong AI so defined. I have tried to demonstrate that the program by itself is not constitutive of thinking because the program is purely a matter of formal symbol manipulation—and we know independently that symbol manipulations by themselves are not sufficient to guarantee the presence of mean-



78

I satisfy the Turing test for understanding Chinese



Computer programs are formal (syntactic).
Human minds have mental contents (semantics)

ings. That is the principle on which the Chinese room argument works.

I emphasize these points here partly because it seems to me the Churchlands [see "Could a Machine Think?" by Paul M. Churchland and Patricia Smith Churchland, page 32] have not quite understood the issues. They think that strong AI is claiming that computers might turn out to think and that I am denying this possibility on commonsense grounds. But that is not the claim of strong AI, and my argument against it has nothing to do with common sense.

I will have more to say about their objections later. Meanwhile I should point out that, contrary to what the Churchlands suggest, the Chinese room argument also refutes any strong-AI claims made for the new parallel technologies that are inspired by and modeled on neural networks. Unlike the traditional von Neumann computer, which proceeds in a step-by-step fashion, these systems have many computational elements that operate in parallel and interact with one another according to rules inspired by neurobiology. Although the results are still modest, these "parallel distributed processing," or "connectionist," models raise useful questions about how complex, parallel network systems like those in brains might actually function in the production of intelligent behavior.

The parallel, "brainlike" character of the processing, however, is irrelevant to the purely computational aspects of the process. Any function that can be computed on a parallel machine can also be computed on a serial machine. Indeed, because parallel machines are still rare, connectionist programs are usually run on traditional serial machines. Parallel processing, then, does not afford a way around the Chinese room argument.

What is more, the connectionist system is subject even on its own terms to a variant of the objection presented by the original Chinese room argument. Imagine that instead of a Chinese room, I have a Chinese gym: a hall containing many monolingual, English-speaking men. These men would carry out the same operations as the nodes and synapses in a connectionist architecture as described by the Churchlands, and the outcome would be the same as having one man manipulate symbols according to a rule book. No one in the gym speaks a word of Chinese, and there is no way for the system as a whole to learn the meanings of any Chinese words. Yet with appropriate adjustments, the system could give the correct answers to Chinese questions.

There are, as I suggested earlier, interesting properties of connectionist nets that enable them to simulate brain processes more accurately than traditional serial architecture does. But the advantages of parallel architecture for weak AI are quite irrelevant to the issues between the Chinese room argument and strong AI.

The Churchlands miss this point when they say that a big enough Chinese gym might have higher-level mental features that emerge from the size and complexity of the system, just as whole brains have mental features that are not had by individual neurons. That is, of course, a possibility, but it has nothing to do with computation. Computationally, serial and parallel systems are equivalent: any computation that can be done in parallel can be done in serial. If the man in the Chinese room is computationally equivalent to both, then if he does not understand Chinese solely by virtue of doing the computations, neither do they. The Churchlands are correct in saying that the original Chinese room argument was designed with traditional AI in mind but wrong in thinking that connectionism is immune to the argument. It applies to any computational system. You can't get semantically loaded thought contents from formal computations alone, whether they are done in serial or in parallel; that is why the Chinese room argument refutes strong AI in any form.

Many people who are impressed by this argument are nonetheless puzzled about the differences between people and computers. If humans are, at least in a trivial sense, computers, and if humans have a semantics, then why couldn't we give semantics to other computers? Why couldn't we program a Vax or a Cray so that it too would have thoughts and feelings? Or why couldn't some new computer technology overcome the gulf between form and content, between syntax and semantics? What, in fact, are the differences between animal brains and computer systems that enable the Chinese room argument to work against computers but not against brains?

The most obvious difference is that the processes that define something as a computer—computational processes—are completely independent of any reference to a specific type of hardware implementation. One could in principle make a computer out of old beer cans strung together with wires and powered by windmills.

But when it comes to brains, although science is largely ignorant of how brains function to produce mental states, one is struck by the extreme specificity of the anatomy and the physiology. Where some understanding exists of how brain processes produce mental phenomena—for example, pain, thirst, vision, smell—it is clear that specific neurobiological processes are involved. Thirst, at least of certain kinds, is caused by certain types of neuron firings in the hypothalamus, which in turn are caused by the action of a specific peptide, angiotensin II. The causation is from the "bottom up" in the sense that lower-level neuronal processes cause higher-level mental phenomena. Indeed, as far as we know, every "mental" event, ranging from feelings of thirst to thoughts of mathematical theorems and memories of childhood, is caused by specific neurons firing in specific neural architectures.

But why should this specificity matter? After all, neuron firings could be simulated on computers that had a completely different physics and chemistry from that of the brain. The answer is that the brain does not merely instantiate a formal pattern or program (it does that, too), but it also *causes* mental events by virtue of specific neurobiological processes. Brains are specific biological organs, and their specific biochemical properties enable them to cause consciousness and other sorts of mental phenomena. Computer simulations of brain processes provide models of the formal aspects of these processes. But the simulation should not be confused with duplication. The computational model of mental processes is no more real than the computational model of any other natural phenomenon.

One can imagine a computer simulation of the action of peptides in the hypothalamus that is accurate down to the last synapse. But equally one can imagine a computer simulation of the oxidation of hydrocarbons in a car engine or the action of digestive processes in a stomach when it is digesting pizza. And the simulation is no more the real thing in the case of the brain than it is in the case of the car or the stomach. Barring miracles, you could not run your car by doing a computer simulation of the oxidation of gasoline, and you could not digest pizza by running the program that simulates such digestion. It seems obvious that a simulation of cognition will similarly not produce the effects of the neurobiology of cognition.

All mental phenomena, then, are caused by neurophysiological processes in the brain. Hence,

Axiom 4. *Brains cause minds.*

In conjunction with my earlier derivation, I immediately derive, trivially,

Conclusion 2. *Any other system capable of causing minds would have to have causal powers (at least) equivalent to those of brains.*

This is like saying that if an electrical engine is to be able to run a car as fast as a gas engine, it must have (at least) an equivalent power output. This conclusion says nothing about the mechanisms. As a matter of fact, cognition is a biological phenomenon: mental states and processes are caused by brain processes. This does not imply that only a biological system could think, but it does imply that any alternative system, whether made of silicon, beer cans or whatever, would have to have the relevant causal capacities equivalent to those of brains. So now I can derive

Conclusion 3. *Any artifact that produced mental phenomena, any artificial brain, would have to be able to duplicate the specific causal powers of brains, and it could not do that just by running a formal program.*

Furthermore, I can derive an important conclusion about human brains:

Conclusion 4. *The way that human brains actually produce mental phenomena cannot be solely by virtue of running a computer program.*

I first presented the Chinese room parable in the pages of *Behavioral and Brain Sciences* in 1980, where it appeared, as is the practice of the journal, along with peer commentary, in this case, 26 commentaries. Frankly, I think the point it makes is rather obvious, but to my surprise the publication was followed by a further flood of objections that—more surprisingly—continues to the present day. The Chinese room argument clearly touched some sensitive nerve.

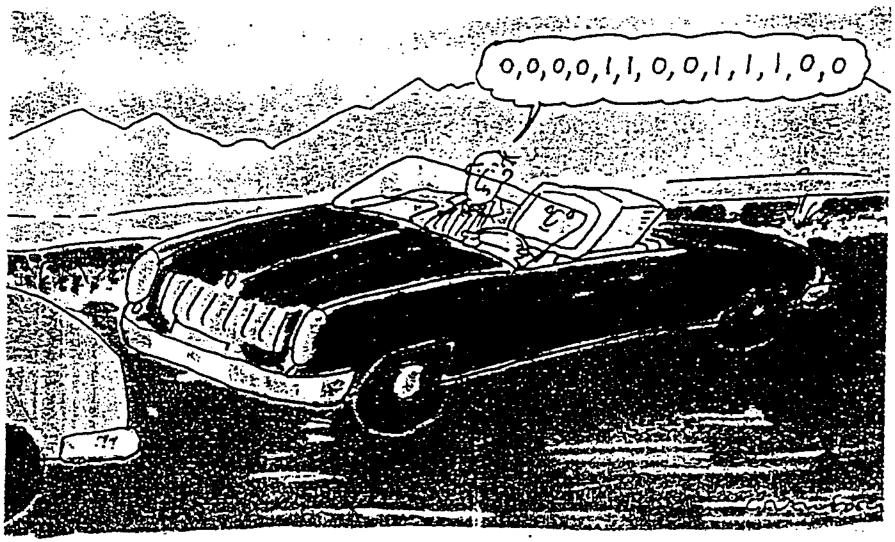
The thesis of strong AI is that any system whatsoever—whether it is made of beer cans, silicon chips or toilet paper—not only might have thoughts and feelings but *must* have thoughts and feelings, provided only that it implements the right program, with the right inputs and outputs. Now, that is a profoundly antibiological view, and one would think that people in AI would be glad to abandon it. Many of them, especially the younger generation, agree with me, but I am amazed at the number and vehemence of the defenders. Here are some of the common objections.

a. In the Chinese room you really do understand Chinese, even though you don't know it. It is, after all, possible to understand something without knowing that one understands it.

b. You don't understand Chinese, but there is an (unconscious) subsystem in you that does. It is, after all, possible to have unconscious mental states, and there is no reason why your understanding of Chinese should not be wholly unconscious.

c. You don't understand Chinese, but the whole room does. You are like a single neuron in the brain, and just as such a single neuron by itself cannot understand but only contributes to the understanding of the whole system, you don't understand, but the whole system does.

d. Semantics doesn't exist anyway; there is only syntax. It is a kind of prescientific illusion to suppose that there exist in the brain some mysterious "mental contents," "thought processes" or "semantics." All that exists in the brain is the same sort of syntactic symbol manipulation that



80

Which semantics is the system giving off now?

goes on in computers. Nothing more.

e. You are not really running the computer program—you only think you are. Once you have a conscious agent going through the steps of the program, it ceases to be a case of implementing a program at all.

f. Computers would have semantics and not just syntax if their inputs and outputs were put in appropriate causal relation to the rest of the world. Imagine that we put the computer into a robot, attached television cameras to the robot's head, installed transducers connecting the television messages to the computer and had the computer output operate the robot's arms and legs. Then the whole system would have a semantics.

g. If the program simulated the operation of the brain of a Chinese speaker, then it would understand Chinese. Suppose that we simulated the brain of a Chinese person at the level of neurons. Then surely such a system would understand Chinese as well as any Chinese person's brain.

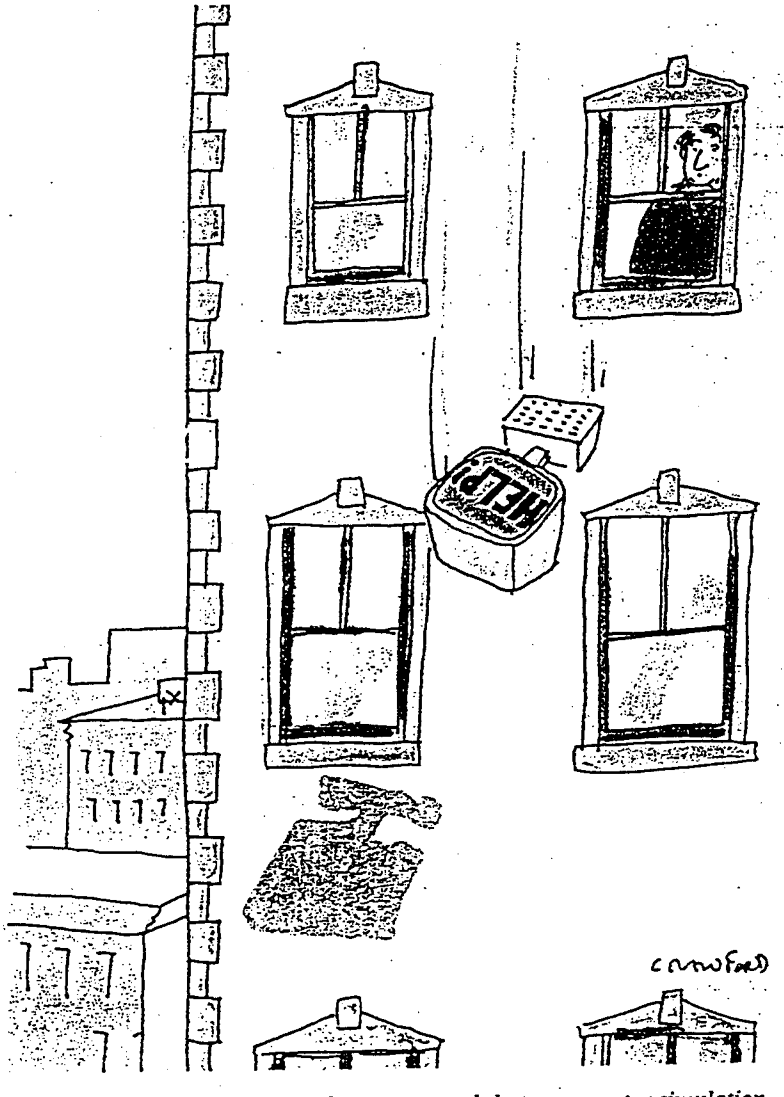
And so on.

All of these arguments share a common feature: they are all inadequate because they fail to come to grips with the actual Chinese room argument. That argument rests on the distinction between the formal symbol manipulation that is done by the computer and the mental contents biologically produced by the brain, a distinction I have abbreviated—I hope not misleadingly—as the distinction between syntax and semantics. I will not repeat my answers to all of these objections, but it will help to clarify the issues if I explain the weaknesses of the most widely held objection, argument c—what I call the systems reply. (The brain simulator reply, argument g, is another popular one, but I have already addressed that one in the previous section.)

The systems reply asserts that of course *you* don't understand Chinese but the whole system—you, the room, the rule book, the bushel baskets full of symbols—does. When I first heard this explanation, I asked one of its proponents, "Do you mean the room understands Chinese?" His answer was yes. It is a daring move, but aside from its implausibility, it will not work on purely logical grounds. The point of the original argument was that symbol shuffling by itself does not give any access to the meanings of the symbols. But this is as much true of the whole room as it is of the person inside. One can see this point by extending the thought experiment. Imagine that I memorize the contents of the baskets and the rule book, and I do all the calculations in my head. You can even imagine that I work out in the open. There is nothing in the "system" that is not in me, and since I don't understand Chinese, neither does the system.

The Churchlands in their companion piece produce a variant of the systems reply by imagining an amusing analogy. Suppose that someone said that light could not be electromagnetic because if you shake a bar magnet in a dark room, the system still will not give off visible light. Now, the Churchlands ask, is not the Chinese room argument just like that? Does it not merely say that if you shake Chinese symbols in a semantically dark room, they will not give off the light of Chinese understanding? But just as later investigation showed that light was entirely constituted by electromagnetic radiation, could not later investigation also show that semantics are entirely constituted of syntax? Is this not a question for further scientific investigation?

Arguments from analogy are notoriously weak, because before one can make the argument work, one has to establish that the two cases are truly analogous. And here I think they are not. The account of light in terms of electromagnetic radiation is a causal story right down to the ground. It is a causal account of the physics of electromagnetic radiation. But the analogy with formal symbols fails because formal symbols have no physical, causal powers. The only power that symbols have, qua symbols, is the power to cause the next step in the program when the machine is running. And there is no question of waiting on further research to reveal the physical,



How could anyone have supposed that a computer simulation of a mental process must be the real thing?

causal properties of 0's and 1's. The only relevant properties of 0's and 1's are abstract computational properties, and they are already well known.

The Churchlands complain that I am "begging the question" when I say that uninterpreted formal symbols are not identical to mental contents. Well, I certainly did not spend much time arguing for it, because I take it as a logical truth. As with any logical truth, one can quickly see that it is true, because one gets inconsistencies if one tries to imagine the converse. So let us try it. Suppose that in the Chinese room some undetectable Chinese thinking really is going on. What exactly is supposed to make the manipulation of the syntactic elements into specifically Chinese thought contents? Well, after all, I am assuming that the programmers were Chinese speakers, programming the system to process Chinese information.

Fine. But now imagine that as I am sitting in the Chinese room shuffling the Chinese symbols, I get bored with just shuffling the—to me—meaningless symbols. So, suppose that I decide to interpret the symbols as standing for moves in a chess game. Which semantics is the system giving off now? Is it giving off a Chinese semantics or a chess semantics, or both simultaneously? Suppose there is a third person looking in through the window, and she decides that the symbol manipulations can all be interpreted as stock-market predictions. And so on. There is no limit to the number of semantic interpretations that can be assigned to the symbols because, to repeat, the symbols are purely formal. They have no intrinsic semantics.

Is there any way to rescue the Churchlands' analogy from incoherence? I said above that formal symbols do not have causal properties. But of course the program will always be implemented in some hardware or another, and the hardware will have specific physical, causal powers. And any real computer will give off various phenomena. My computers, for example, give off heat, and they make a humming noise and sometimes crunching sounds. So is there some logically compelling reason why they could not also give off consciousness? No. Scientifically, the idea is out of the question, but it is not something the Chinese room argument is supposed to refute, and it is not something that an adherent of strong AI would wish to defend, because any such giving off would have to derive from the physical features of the implementing medium. But the basic premise of strong AI is that the physical features of the implementing medium are totally irrelevant. What matters are programs, and programs are purely formal.

The Churchlands' analogy between syntax and electromagnetism, then, is confronted with a dilemma; either the syntax is construed purely formally in terms of its abstract mathematical properties, or it is not. If it is, then the analogy breaks down, because syntax so construed has no physical powers and hence no physical, causal powers. If, on the other hand, one is supposed to think in terms of the physics of the implementing medium, then there is indeed an analogy, but it is not one that is relevant to strong AI.

Because the points I have been making are rather obvious—syntax is not the same as semantics, brain processes cause mental phenomena—the question arises, How did we get into this mess? How could anyone have supposed that a computer simulation of a mental process must be the real thing? After all, the whole point of models is that they contain only certain features of the modeled domain and leave out the rest. No one expects to get wet in a pool filled with Ping-Pong-ball models of water molecules. So why would anyone think a computer model of thought processes would actually think?

Part of the answer is that people have inherited a residue of behaviorist psychological theories of the past generation. [The Turing test enshrines the temptation to think that if something behaves as if it had certain mental processes, then it must actually have those mental processes.] And this is part of the behaviorists' mistaken assumption that in order to be scientific, psychology must confine its study to externally observable behavior. Paradoxically, this residual behaviorism is tied to a residual dualism. Nobody thinks that a computer simulation of digestion would actually digest anything, but where cognition is concerned, people are willing to believe in such a miracle because they fail to recognize that the mind is just as much a biological phenomenon as digestion. The mind, they suppose, is something formal and abstract, not a part of the wet and slimy stuff in our heads. The polemical literature in AI usually contains attacks on something the authors call dualism, but what they fail to see is that they themselves display dualism in a strong form, for unless one accepts the idea that the mind is completely independent of the brain or of any other physically specific system, one could not possibly hope to create minds just by designing programs.

Historically, scientific developments in the West that have treated humans as just a part of the ordinary physical, biological order have often been opposed by various rearguard actions. Copernicus and Galileo were opposed because they denied that the earth was the center of the universe; Darwin was opposed because he claimed that humans had descended from the lower animals. It is best to see strong AI as one of the last gasps of this antiscientific tradition, for it denies that there is anything essentially physical and biological about the human mind. The mind according to strong AI is independent of the brain. It is a computer program and as such has no essential connection to any specific hardware.

Many people who have doubts about the psychological significance of AI think that computers might be able to understand Chinese and think about numbers but cannot do the crucially human things, namely—and then follows their favorite human specialty—falling in love, having a sense of humor, feeling the angst of postindustrial society under late capitalism, or whatever. But workers in AI complain—correctly—that this is a case of moving the goalposts. As soon as an AI simulation succeeds, it ceases to be of psychological importance. In this debate both sides fail to see the distinction between simulation and duplication. As far as simulation is concerned, there is no difficulty in programming my computer so that it prints out, "I love you, Suzy"; "Ha ha"; or "I am suffering the angst of postindustrial society under late capitalism." The important point is that simulation is not the same as duplication, and that fact holds as much import for thinking about arithmetic as it does for feeling angst. The point is not that the computer gets only to the 40-yard line and not all the way to the goal line. The computer doesn't even get started. It is not playing that game.

FURTHER READING

MIND DESIGN: PHILOSOPHY, PSYCHOLOGY, ARTIFICIAL INTELLIGENCE. Edited by John Haugeland. The MIT Press, 1980.

MINDS, BRAINS, AND PROGRAMS. John Searle in *Behavioral and Brain Sciences*, Vol. 3, No. 3, pages 417–458; 1980.

MINDS, BRAINS, AND SCIENCE. John R. Searle. Harvard University Press, 1984.

MINDS, MACHINES AND SEARLE. Stevan Harnad in *Journal of Experimental and Theoretical Artificial Intelligence*, Vol. 1, No. 1, pages 5–25; 1989.

Could a Machine Think?

Classical AI is unlikely to yield conscious machines; systems that mimic the brain might

by Paul M. Churchland and Patricia Smith Churchland

Artificial-intelligence research is undergoing a revolution. To explain how and why, and to put John R. Searle's argument in perspective, we first need a flashback.

By the early 1950's the old, vague question, Could a machine think? had been replaced by the more approachable question, Could a machine that manipulated physical symbols according to structure-sensitive rules think? This question was an improvement because formal logic and computational theory had seen major developments in the preceding half-century. Theorists had come to appreciate the enormous power of abstract systems of symbols that undergo rule-governed transformations. If those systems could just be automated, then their abstract computational power, it seemed, would be displayed in a real physical system. This insight spawned a well-defined research program with deep theoretical underpinnings.

Could a machine think? There were many reasons for saying yes. One of the earliest and deepest reasons lay in two important results in computational theory. The first was Church's thesis, which states that every effectively computable function is recursively computable. Effectively computable means that there is a "rote" procedure for determining, in finite time, the output of the function for a given input. Recursively computable means more specifically that there is a finite set of operations that can be applied to a given input, and then applied again and again to the successive results of such applications, to yield the function's output in finite time. The notion of a rote procedure is nonformal and intuitive; thus, Church's thesis does not admit of a formal proof. But it does go to the heart of what it is to compute, and many lines of evidence converge in supporting it.

PAUL M. CHURCHLAND and PATRICIA SMITH CHURCHLAND are professors of philosophy at the University of California at San Diego. Together they have studied the nature of the mind and knowledge for the past two decades. Paul Churchland focuses on the nature of scientific knowledge and its development, while Patricia Churchland focuses on the neurosciences and on how the brain sustains cognition. Paul Churchland's *Matter and Consciousness* is the standard textbook on the philosophy of the mind, and Patricia Churchland's *Neurophilosophy* brings together theories of cognition from both philosophy and biology. Paul Churchland is currently chair of the philosophy department at UCSD, and the two are, respectively, president and past president of the Society for Philosophy and Psychology. Patricia Churchland is also an adjunct professor at the Salk Institute for Biological Studies in San Diego. The Churchlands are also members of the UCSD cognitive science faculty, its Institute for Neural Computation and its Science Studies program.

The second important result was Alan M. Turing's demonstration that any recursively computable function can be computed in finite time by a maximally simple sort of symbol-manipulating machine that has come to be called a universal Turing machine. This machine is guided by a set of recursively applicable rules that are sensitive to the identity, order and arrangement of the elementary symbols it encounters as input.

These two results entail something remarkable, namely that a standard digital computer, given only the right program, a large enough memory and sufficient time, can compute *any* rule-governed input-output function. That is, it can display any systematic pattern of responses to the environment whatsoever.

More specifically, these results imply that a suitably programmed symbol-manipulating machine (hereafter, SM machine) should be able to pass the Turing test for conscious intelligence. The Turing test is a purely behavioral test for conscious intelligence, but it is a very demanding test even so. (Whether it is a fair test will be addressed below, where we shall also encounter a second and quite different "test" for conscious intelligence.) In the original version of the Turing test, the inputs to the SM machine are conversational questions and remarks typed into a console by you or me, and the outputs are typewritten responses from the SM machine. The machine passes this test for conscious intelligence if its responses cannot be discriminated from the typewritten responses of a real, intelligent person. Of course, at present no one knows the function that would produce the output behavior of a conscious person. But the Church and Turing results assure us that, whatever that (presumably effective) function might be, a suitable SM machine could compute it.

This is a significant conclusion, especially since Turing's portrayal of a purely teletyped interaction is an unnecessary restriction. The same conclusion follows even if the SM machine interacts with the world in more complex ways: by direct vision, real speech and so forth. After all, a more complex recursive function is still Turing-computable. The only remaining problem is to identify the undoubtedly complex function that governs the human pattern of response to the environment and then write the program (the set of recursively applicable rules) by which the SM machine will compute it. These goals form the fundamental research program of classical AI.

Initial results were positive. SM machines with clever programs performed a variety of ostensibly cognitive activities. They responded to complex instructions, solved complex arithmetic, algebraic and tactical problems, played checkers and chess, proved theorems and engaged in simple dialogue. Performance continued to improve with the appearance of larger memories and faster machines and with the use of longer and more cunning programs. Classical, or "program-writing," AI was a vigorous and successful research effort from almost every perspective. The occasional denial that an SM machine might eventually think appeared uninformed and ill motivated. The case for a positive answer to our title question was overwhelming.

There were a few puzzles, of course. For one thing, SM machines were admittedly not very brainlike. Even here, however, the classical approach had a convincing answer. First, the physical material of any SM machine has nothing essential to do with what function it computes. That is fixed by its program. Second, the engineering details of any machine's functional architecture are also irrelevant, since different architectures running quite different programs can still be computing the same input-output function.

Accordingly, AI sought to find the input-output *function* characteristic of intelligence and the most efficient of the many possible programs for computing it. The idiosyncratic way in which the brain computes the function just doesn't matter, it was said. This completes the rationale for classical AI and for a positive answer to our title question.

Could a machine think? There were also some arguments for saying no. Through the 1960's interesting negative arguments were relatively rare. The objection was occasionally made that thinking was a nonphysical process in an immaterial soul. But such dualistic resistance was neither evolutionarily nor explanatorily plausible. It had a negligible impact on AI research.

A quite different line of objection was more successful in gaining the AI community's attention. In 1972 Hubert L. Dreyfus published a book that was highly critical of the parade-case simulations of cognitive activity. He argued for their inadequacy as simulations of genuine cognition, and he pointed to a pattern of failure in these attempts. What they were missing, he suggested, was the vast store of inarticulate background knowledge every person possesses and the commonsense capacity for drawing on relevant aspects of that knowledge as changing circumstance demands. Dreyfus did not deny the possibility that an artificial physical system of some kind might think, but he was highly critical of the idea that this could be achieved solely by symbol manipulation at the hands of recursively applicable rules.

Dreyfus's complaints were broadly perceived within the AI community, and within the discipline of philosophy as well, as shortsighted and unsympathetic, as harping on the inevitable simplifications of a research effort still in its youth. These deficits might be real, but surely they were temporary. Bigger machines and better programs should repair them in due course. Time, it was felt, was on AI's side. Here again the impact on research was negligible.

Time was on Dreyfus's side as well: the rate of cognitive return on increasing speed and memory began to slacken in the late 1970's and early 1980's. The simulation of object recognition in the visual system, for example, proved computationally intensive to an unexpected degree. Realistic results required longer and longer periods of computer time, periods far in excess of what a real visual system requires. This relative slowness of the simulations was darkly curious; signal propagation in a computer is roughly a million times faster than in the brain, and the clock frequency of a computer's central processor is greater than any frequency found in the brain by a similarly dramatic margin. And yet, on realistic problems, the tortoise easily outran the hare.

Furthermore, realistic performance required that the computer program have access to an extremely large knowledge base. Constructing the relevant knowledge base was problem enough, and it was compounded by the problem of how to access just the contextually relevant parts of that knowledge base in real time. As the knowledge base got bigger and better, the access problem got worse. Exhaustive search took too much time, and heuristics for relevance did poorly. Worries of the sort Dreyfus had raised finally began to take hold here

THE CHINESE ROOM

Axiom 1. Computer programs are formal (syntactic).

Axiom 2. Human minds have mental contents (semantics).

Axiom 3. Syntax by itself is neither constitutive of nor sufficient for semantics.

Conclusion 1. Programs are neither constitutive of nor sufficient for minds.

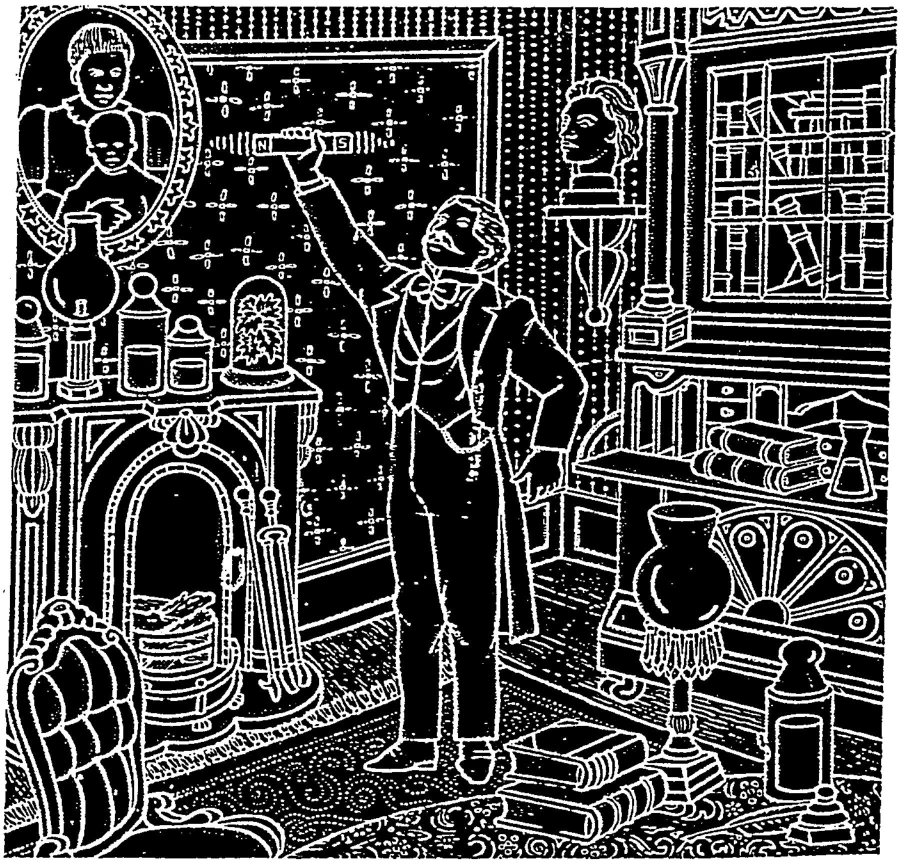
THE LUMINOUS ROOM

Axiom 1. Electricity and magnetism are forces.

Axiom 2. The essential property of light is luminance.

Axiom 3. Forces by themselves are neither constitutive of nor sufficient for luminance.

Conclusion 1. Electricity and magnetism are neither constitutive of nor sufficient for light.



OSCILLATING ELECTROMAGNETIC FORCES constitute light even though a magnet pumped by a person appears to produce no light whatsoever. Similarly, rule-based symbol manipulation might constitute intelligence even though the rule-based system inside John R. Searle's "Chinese room" appears to lack real understanding.

and there even among AI researchers.

At about this time (1980) John Searle authored a new and quite different criticism aimed at the most basic assumption of the classical research program: the idea that the appropriate manipulation of structured symbols by the recursive application of structure-sensitive rules could constitute conscious intelligence.

Searle's argument is based on a thought experiment that displays two crucial features. First, he describes a SM machine that realizes, we are to suppose, an input-output function adequate to sustain a successful Turing test conversation conducted entirely in Chinese. Second, the internal structure of the machine is such that, however it behaves, an observer remains certain that neither the machine nor any part of it understands Chinese. All it contains is a monolingual English speaker following a written set of instructions for manipulating the Chinese symbols that arrive and leave through a mail slot. In short, the system is supposed to pass the Turing test, while the system itself lacks any genuine understanding of Chinese or real Chinese semantic content [see "Is the Brain's Mind a Computer Program?" by John R. Searle, page 26].

The general lesson drawn is that any system that merely manipulates physical symbols in accordance with structure-sensitive rules will be at best a hollow mock-up of real conscious intelligence, because it is impossible to generate "real semantics" merely by cranking away on "empty syntax." Here, we should point out, Searle is imposing a nonbehavioral test for consciousness: the elements of conscious intelligence must possess real semantic content.

One is tempted to complain that Searle's thought experiment is unfair because his Rube Goldberg system will compute with absurd slowness. Searle insists, however, that speed is strictly irrelevant here. A slow thinker should still be a real thinker. Everything essential to the duplication of thought, as per classical AI, is said to be present in the Chinese room.

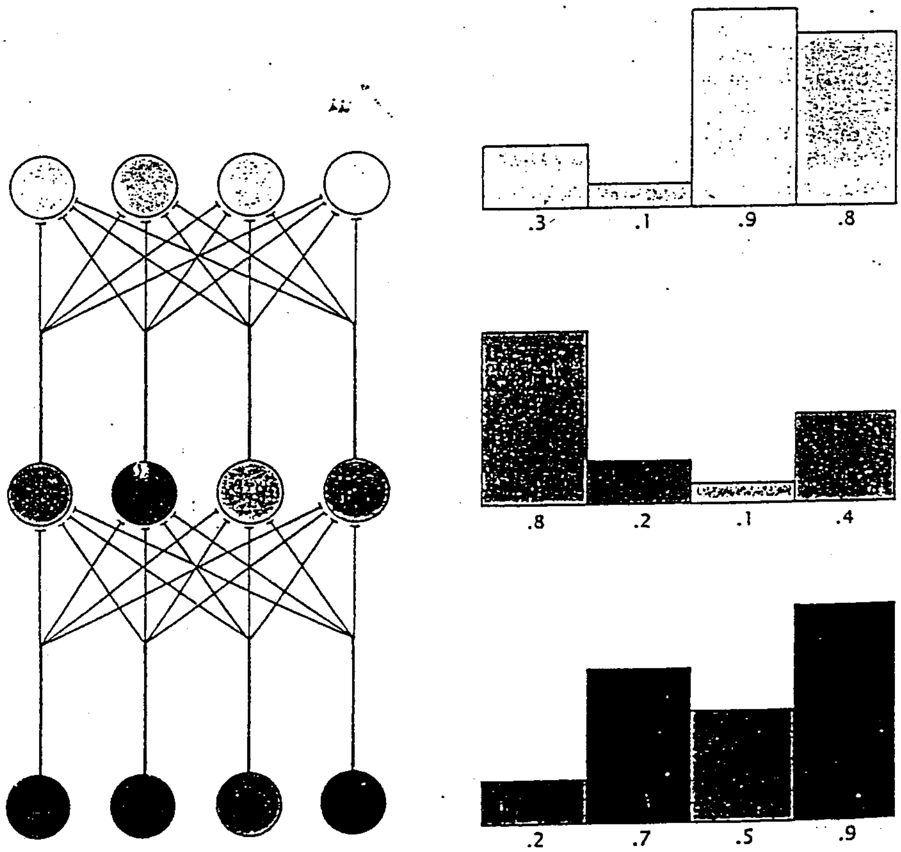
Searle's paper provoked a lively reaction from AI researchers, psychologists and philosophers alike. On the whole, however, he was met with an even more hostile reception than Dreyfus had experienced. In his companion piece in this issue, Searle forthrightly lists a number of these critical responses. We think many of them are reasonable, especially those that "bite the bullet" by insisting that, although it is appallingly slow, the overall system of the room-plus-contents does understand Chinese.

We think those are good responses, but not because we think that the room understands Chinese. We agree with Searle that it does not. Rather they are good responses because they reflect a refusal to accept the crucial third axiom of Searle's argument: "*Syntax by itself is neither constitutive of nor sufficient for semantics.*" Perhaps this axiom is true, but Searle cannot rightly pretend to know that it is. Moreover, to assume its truth is tantamount to begging the question against the research program of classical AI, for that program is predicated on the very interesting assumption that if one can just set in motion an appropriately structured internal dance of syntactic elements, appropriately connected to inputs and outputs, it can produce the same cognitive states and achievements found in human beings.

The question-begging character of Searle's axiom 3 becomes clear when it is compared directly with his conclusion 1: "*Programs are neither constitutive of nor sufficient for minds.*" Plainly, his third axiom is already carrying 90 percent of the weight of this almost identical conclusion. That is why Searle's thought experiment is devoted to shoring up axiom 3 specifically. That is the point of the Chinese room.

Although the story of the Chinese room makes axiom 3 tempting to the unwary, we do not think it succeeds in establishing axiom 3, and we offer a parallel argument below in illustration of its failure. A single transparently fallacious instance of a disputed argument often provides far more insight than a book full of logic chopping.

Searle's style of skepticism has ample precedent in the history of science. The 18th-century Irish bishop George Berkeley found it unintelligible that compression waves in the air, by themselves, could constitute or be sufficient for objective sound. The English poet-artist William Blake and the German poet-naturalist Johann W.



NEURAL NETWORKS model a central feature of the brain's microstructure. In this three-layer net, input neurons (*bottom left*) process a pattern of activations (*bottom right*) and pass it along weighted connections to a hidden layer. Elements in the hidden layer sum their many inputs to produce a new pattern of activations. This is passed to the output layer, which performs a further transformation. Overall the network transforms any input pattern into a corresponding output pattern as dictated by the arrangement and strength of the many connections between neurons.

85

von Goethe found it inconceivable that small particles by themselves could constitute or be sufficient for the objective phenomenon of light. Even in this century, there have been people who found it beyond imagining that inanimate matter by itself, and however organized, could ever constitute or be sufficient for life. Plainly, what people can or cannot imagine often has nothing to do with what is or is not the case, even where the people involved are highly intelligent.

To see how this lesson applies to Searle's case, consider a deliberately manufactured parallel to his argument and its supporting thought experiment.

Axiom 1. *Electricity and magnetism are forces.*

Axiom 2. *The essential property of light is luminance.*

Axiom 3. *Forces by themselves are neither constitutive of nor sufficient for luminance.*

Conclusion 1. *Electricity and magnetism are neither constitutive of nor sufficient for light.*

Imagine this argument raised shortly after James Clerk Maxwell's 1864 suggestion that light and electromagnetic waves are identical but before the world's full appreciation of the systematic parallels between the properties of light and the properties of electromagnetic waves. This argument could have served as a compelling objection to Maxwell's imaginative hypothesis, especially if it were accompanied by the following commentary in support of axiom 3.

"Consider a dark room containing a man holding a bar magnet or charged object. If the man pumps the magnet up and down, then, according to Maxwell's theory of artificial luminance (AL), it will initiate a spreading circle of electromagnetic waves and will thus be luminous. But as all of us who have toyed with magnets or charged balls well know, their forces (or any other forces for that matter), even when set in motion, produce no luminance at all. It is inconceivable that you might constitute real luminance just by moving forces around!"

How should Maxwell respond to this challenge? He might begin by insisting that the "luminous room" experiment is a misleading display of the phenomenon of luminance because the frequency of oscillation of the magnet is absurdly low, too low by a factor of 10^{15}. This might well elicit the impatient response that frequency has nothing to do with it, that the room with the bobbing magnet already contains everything essential to light, according to Maxwell's own theory.

In response Maxwell might bite the bullet and claim, quite correctly, that the room really is bathed in luminance, albeit a grade or quality too feeble to appreciate. (Given the low frequency with which the man can oscillate the magnet, the wavelength of the electromagnetic waves produced is far too long and their intensity is much too weak for human retinas to respond to them.) But in the climate of understanding here contemplated—the 1860's—this tactic is likely to elicit laughter and hoots of derision. "Luminous room, my foot, Mr. Maxwell. It's pitch-black in there!"

Alas, poor Maxwell has no easy route out of this predicament. All he can do is insist on the following three points. First, axiom 3 of the above argument is false. Indeed, it begs the question despite its intuitive plausibility. Second, the luminous room experiment demonstrates nothing of interest one way or the other about the nature of light. And third, what is needed to settle the problem of light and the possibility of artificial luminance is an ongoing research program to determine whether under the appropriate conditions the behavior of electromagnetic waves does indeed mirror perfectly the behavior of light.

This is also the response that classical AI should give to Searle's argument. Even though Searle's Chinese room may appear to be "semantically dark," he is in no position to insist, on the strength of this appearance, that rule-governed symbol manipulation can never constitute semantic phenomena, especially when people have only an uninformed common-sense understanding of the semantic and cognitive phenomena that need to be explained. Rather than exploit one's understanding of these things, Searle's argument freely exploits one's ignorance of them.

With these criticisms of Searle's argument in place, we return to the question of whether the research program of classical AI has a realistic chance of solving the problem of conscious intelligence and of producing a machine that thinks. We believe that the prospects are poor, but we rest this opinion on reasons very different from Searle's. Our reasons derive from the specific performance failures of the classical research program in AI and from a variety of lessons learned from the biological brain and a new class of computational models inspired by its structure. We have already indicated some of the failures of classical AI regarding tasks that the brain performs swiftly and efficiently. The emerging consensus on these failures is that the functional architecture of classical SM machines is simply the wrong architecture for the very demanding jobs required.

What we need to know is this: How does the brain achieve cognition? Reverse engineering is a common practice in industry. When a new piece of technology comes on the market, competitors find out how it works by taking it apart and divining its structural rationale. In the case of the brain, this strategy presents an unusually stiff challenge, for the brain is the most complicated and sophisticated thing on the planet. Even so, the neurosciences have revealed much about the brain on a wide variety of structural levels. Three anatomic points will provide a basic contrast with the architecture of conventional electronic computers.

First, nervous systems are parallel machines, in the sense that signals are processed in millions of different pathways simultaneously. The retina, for example, presents its complex input to the brain not in chunks of eight, 16 or 32 elements, as in a desktop computer, but rather in the form of almost a million distinct signal elements arriving simultaneously at the target of the optic nerve (the lateral geniculate nucleus), there to be processed collectively, simultaneously and in one fell swoop. Second, the brain's basic processing unit, the neuron, is comparatively simple. Furthermore, its response to incoming signals is analog, not digital, inasmuch as its output spiking frequency varies continuously with its input signals. Third, in the brain, axons projecting from one neuronal population to another are often matched by axons returning from their target population. These descending or recurrent projections allow the brain to modulate the character of its sensory processing. More important still, their existence makes the brain a genuine dynamical system whose continuing behavior is both highly complex and to some degree independent of its peripheral stimuli.

Highly simplified model networks have been useful in suggesting how real neural networks might work and in revealing the computational properties of parallel architectures. For example, consider a three-layer model consisting of neuronlike units fully connected by axonlike connections to the units at the next layer. An input stimulus produces some activation level in a given input unit, which conveys a signal of proportional strength along its "axon" to its many "synaptic" connections to the hidden units. The global effect is that a pattern of activations across the set of input units produces a distinct pattern of activations across the set of hidden units.

The same story applies to the output units. As before, an activation pattern across the hidden units produces a distinct activation pattern across the output units. All told, this network is a device for transforming any one of a great many possible input vectors (activation patterns) into a uniquely corresponding output vector. It is a device for computing a specific function. Exactly which function it computes is fixed by the global configuration of its synaptic weights.

There are various procedures for adjusting the weights so as to yield a network that computes almost any function—that is, any vector-to-vector transformation—that one might desire. In fact, one can even impose on it a function one is unable to specify, so long as one can supply a set of examples of the desired input-output pairs. This process, called "training up the network," proceeds by successive adjustment of the network's weights until it performs the input-output transformations desired.

עיקרי PDP:

Although this model network vastly oversimplifies the structure of the brain, it does illustrate several important ideas. First, a parallel architecture provides a dramatic speed advantage over a conventional computer, for the many synapses at each level perform many small computations simultaneously instead of in laborious sequence. This advantage gets larger as the number of neurons increases at each layer. Strikingly, the speed of processing is entirely independent of both the number of units involved in each layer and the complexity of the function they are computing. Each layer could have four units or a hundred million; its configuration of synaptic weights could be computing simple one-digit sums or second-order differential equations. It would make no difference. The computation time would be exactly the same.

[illegible]
חישובים
מקבילים

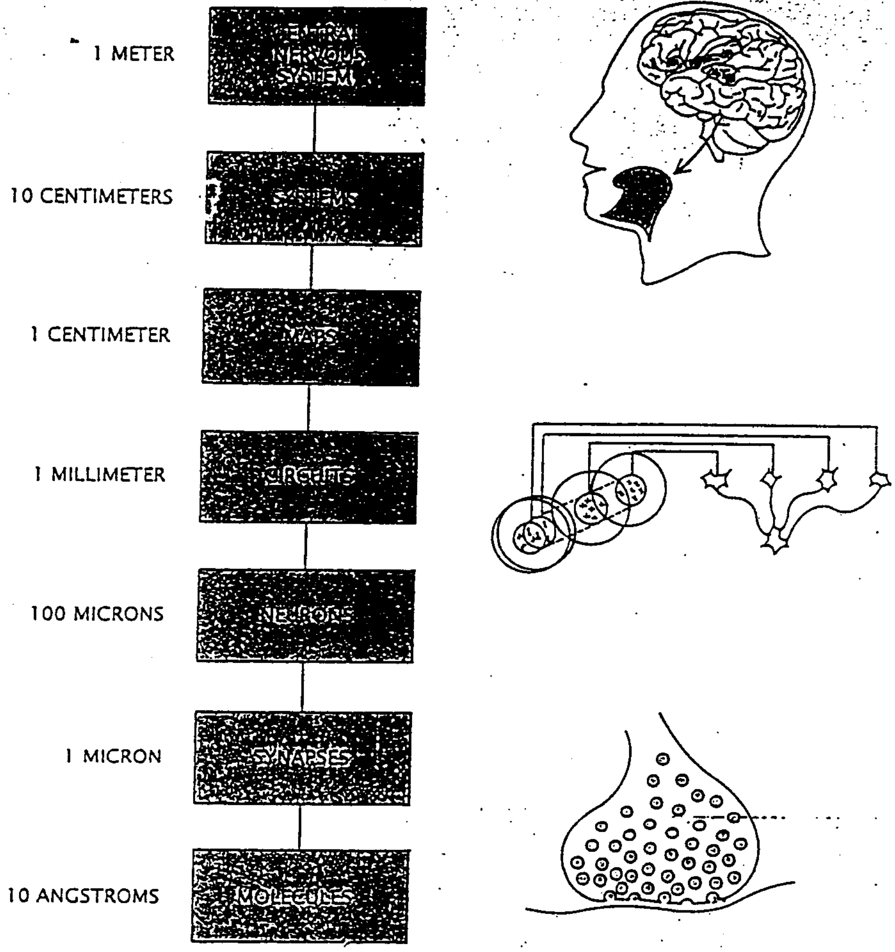
Second, massive parallelism means that the system is fault-tolerant and functionally persistent; the loss of a few connections, even quite a few, has a negligible effect on the character of the overall transformation performed by the surviving network.

Third, a parallel system stores large amounts of information in a distributed fashion, any part of which can be accessed in milliseconds. That information is stored in the specific configuration of synaptic connection strengths, as shaped by past learning. Relevant information is "released" as the input vector passes through—and is transformed by—that configuration of connections.

Parallel processing is not ideal for all types of computation. On tasks that require only a small input vector, but many millions of swiftly iterated recursive computations, the brain performs very badly, whereas classical SM machines excel. This class of computations is very large and important, so classical machines will always be useful, indeed, vital. There is, however, an equally large class of computations for which the brain's architecture is the superior technology. These are the computations that typically confront living creatures: recognizing a predator's outline in a noisy environment; recalling instantly how to avoid its gaze, flee its approach or fend off its attack; distinguishing food from nonfood and mates from nonmates; navigating through a complex and ever-changing physical/social environment; and so on.

Finally, it is important to note that the parallel system described is not manipulating symbols according to structure-sensitive rules. Rather symbol manipulation appears to be just one of many cognitive skills that a network may or may not learn to display. Rule-governed symbol manipulation is not its basic mode of operation. Searle's argument is directed against rule-governed SM machines; vector transformers of the kind we describe are therefore not threatened by his Chinese room argument even if it were sound, which we have found independent reason to doubt.

Searle is aware of parallel processors but thinks they too will be devoid of real semantic content. To illustrate their inevitable failure, he outlines a



NERVOUS SYSTEMS span many scales of organization, from neurotransmitter molecules (*bottom*) to the entire brain and spinal cord. Intermediate levels include single neurons and circuits made up of a few neurons, such as those that produce orientation selectivity to a visual stimulus (*middle*), and systems made up of circuits such as those that subserve language (*top right*). Only research can decide how closely an artificial system must mimic the biological one to be capable of intelligence.

second thought experiment, the Chinese gym, which has a gymnasium full of people organized into a parallel network. From there his argument proceeds as in the Chinese room.

We find this second story far less responsive or compelling than his first. For one, it is irrelevant that no unit in his system understands Chinese, since the same is true of nervous systems: no neuron in my brain understands English, although my whole brain does. For another, Searle neglects to mention that his simulation (using one person per neuron, plus a fleet-footed child for each synaptic connection) will require at least 10^{14} people, since the human brain has 10^{11} neurons, each of which averages over 10^3 connections. His system will require the entire human populations of over 10,000 earths. One gymnasium will not begin to hold a fair simulation.

[illegible]

On the other hand, if such a system were to be assembled on a suitably cosmic scale, with all its pathways faithfully modeled on the human case, we might then have a large, slow, oddly made but still functional brain on our hands. In that case the default assumption is surely that, given proper inputs, it would think, not that it couldn't. There is no guarantee that its activity would constitute real thought, because the vector-processing theory sketched above may not be the correct theory of how brains work. But neither is there any a priori guarantee that it could not be thinking. Searle is once more mistaking the limits on his (or the reader's) current imagination for the limits on objective reality.

The brain is a kind of computer, although most of its properties remain to be discovered. Characterizing the brain as a kind of computer is neither trivial nor frivolous. The brain does compute functions, functions of great complexity, but not in the classical AI fashion. When brains are said to be computers, it should not be implied that they are serial, digital computers, that they are programmed, that they exhibit the distinction between hardware and software or that they must be symbol manipulators or rule followers. Brains are computers in a radically different style.

How the brain manages meaning is still unknown, but it is clear that the problem reaches beyond language use and beyond humans. A small mound of fresh dirt signifies to a person, and also to coyotes, that a gopher is around; an echo with a certain spectral character signifies to a bat the presence of a moth. To develop a theory of meaning, more must be known about how neurons code and transform sensory signals, about the neural basis of memory, learning and emotion and about the interaction of these capacities and the motor system. A neurally grounded theory of meaning may require revision of the very intuitions that now seem so secure and that are so freely exploited in Searle's arguments. Such revisions are common in the history of science.

Could science construct an artificial intelligence by exploiting what is known about the nervous system? We see no principled reason why not. Searle appears to agree, although he qualifies his claim by saying that "any other system capable of causing minds would have to have causal powers (at least) equivalent to those of brains." We close by addressing this claim. We presume that Searle is not claiming that a successful artificial mind must have *all* the causal powers of the brain, such as the power to smell bad when rotting, to harbor slow viruses such as kuru, to stain yellow with horseradish peroxidase and so forth. Requiring perfect parity would be like requiring that an artificial flying device lay eggs.

Presumably he means only to require of an artificial mind all of the causal powers relevant, as he says, to conscious intelligence. But which exactly are they? We are back to quarreling about what is and is not relevant. This is an entirely reasonable place for a disagreement, but it is an empirical matter, to be tried and tested. Because so little is known about what goes into the process of cognition and semantics, it is premature to be very confident about what features are essential. Searle hints at various points that every level, including the biochemical, must be represented in any machine that is a candidate for artificial intelligence. This claim is almost surely too strong. An artificial brain might use something other than biochemicals to achieve the same ends.

This possibility is illustrated by Carver A. Mead's research at the California Institute of Technology. Mead and his colleagues have used analog VLSI techniques to build an artificial retina and an artificial cochlea. (In animals the retina and cochlea are not mere transducers: both systems embody a complex processing network.) These are not mere simulations in a minicomputer of the kind that Searle derides; they are real information-processing units responding in real time to real light, in the case of the artificial retina, and to real sound, in the case of the artificial cochlea. Their circuitry is based on the known anatomy and physiology of the cat retina and the barn owl cochlea, and their output is dramatically similar to the known output of the organs at issue.

These chips do not use any neurochemicals, so neurochemicals are clearly not necessary to achieve the evident results. Of course, the artificial retina cannot be said to see anything, because its output does not have an artificial thalamus or cortex to go to. Whether Mead's program could be sustained to build an entire artificial brain remains to be seen, but there is no evidence now that the absence of biochemicals renders it quixotic.

We, and Searle, reject the Turing test as a sufficient condition for conscious intelligence. At one level our reasons for doing so are similar: we agree that it is also very important how the input-output function is achieved; it is important that the right sorts of things be going on inside the artificial machine. At another level, our reasons are quite different. Searle bases his position on commonsense intuitions about the presence or absence of semantic content. We base ours on the specific behavioral failures of the classical SM machines and on the specific virtues of machines with a more brainlike architecture. These contrasts show that certain computational strategies have vast and decisive advantages over others where typical cognitive tasks are concerned, advantages that are empirically inescapable. Clearly, the brain is making systematic use of these computational advantages. But it need not be the only physical system capable of doing so. Artificial intelligence, in a nonbiological but massively parallel machine, remains a compelling and discernible prospect.

FURTHER READING

COMPUTING MACHINERY AND INTELLIGENCE. Alan M. Turing in *Mind*, Vol. 59, pages 433–460; 1950.

WHAT COMPUTERS CAN'T DO; A CRITIQUE OF ARTIFICIAL REASON. Hubert L. Dreyfus. Harper & Row, 1972.

NEUROPHILOSOPHY: TOWARD A UNIFIED UNDERSTANDING OF THE MIND/BRAIN. Patricia Smith Churchland. The MIT Press, 1986.

FAST THINKING in *The Intentional Stance*. Daniel Clement Dennett. The MIT Press, 1987.

A NEUROCOMPUTATIONAL PERSPECTIVE: THE NATURE OF MIND AND THE STRUCTURE OF SCIENCE. Paul M. Churchland. The MIT Press, in press.

88