24.09x Minds and Machines

John R. Searle, "Is the brain's mind a computer program?"

Excerpts from John R. Searle, "Is the brain's mind a computer program?" (*Scientific American* 262: 26-31, 1990)

Searle begins by distinguishing two sorts of questions. First:

> Can a machine think? Can a machine have conscious thoughts in exactly the same sense that you and I have?

Second:

> 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.

Searle is interested in this answer to the second sort of question:

> 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.

Searle distinguishes between strong and weak AI:

> 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."

He then purports to give a counterexample to strong AI.

> 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 squiggle-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.

He summarizes the point of this thought experiment:

> [I]f 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.

> Searle now puts the argument against strong AI in premise-conclusion form.
>
> 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[1] 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.

Searle’s second premise is:

[1] As (Latin).

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

He explains the second premise as follows:

> [This] 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.

Finally, the third premise:

> 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.

These premises, Searle says, imply:

> Conclusion 1. *Programs are neither constitutive of nor sufficient for minds.*
>
> And that is just another way of saying that strong AI is false.

Searle now clarifies what his argument is supposed to prove, and what it isn't supposed to prove:

> 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 demon-strate 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 meanings. That is the principle on which the Chinese room argument works.

Searle claims that his argument is indifferent to the details of the computer architecture:

> Contrary to what the Churchlands suggest[2], 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.

An adaptation of the Chinese room argument is supposed to show that parallel processing doesn't help rescue strong AI:

> 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

[2] Searle is referring to "Could a machine think?", by Paul M. Churchland and Patricia Smith Churchland", in the same issue of *Scientific American*.

> 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.

Again, he stresses that parallel and non-parallel systems are computationally equivalent.

> 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.

Searle then considers common objections to his argument:

> 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 pre scientific 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 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.

Searle has a general diagnosis of these objections:

> …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.

He only addresses the systems reply (c) at length.

> 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.

He claims that the Churchlands' response is a variant of the systems reply. He describes it as follows:

> 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?

Searle first objects to the argument by saying the analogy is not a good one.

> 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 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.

Searle then responds to the charge of question begging.

> 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.

Searle imagines one way of rescuing the Churchlands' analogy. But he argues this way is irrelevant to what he is trying to refute.

> 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.

According to Searle, the analogy faces a dilemma:

> 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.

Searle concludes:

> 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.
>
> This is the mistake of confusing simulation with 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.