The Ghost is The Machine: A Defense of the

MATT CARLSON

In "Minds, Brains and Science," John Searle attempts to show that the mind is necessarily more than simply an instantiation of a computer program. Parts of Searle's argument are quite persuasive; indeed some of his conclusions are both valid and, I believe, sound. However, his overall conclusion is somewhat misleading. The thesis that Searle refutes (strong artificial intelligence, or AI, as he terms it) is essentially the view that the brain is a sort of hardware and mind is a sort of software. Thus, according to this thesis, if we could program the correct software, we could program a mind. By refuting this thesis, Searle seems to have refuted all claims to the possibility of artificial intelligence. However, I will attempt to show that Searle's argument misses the mark and that, while the mind is not like a program running on certain hardware, the brain is. In so doing, I will also address the adequacy of the Turing test as a test of AI, as Searle's argument does show the worth of this test to be suspect.

The Turing Test and Artificial Intelligence

The Turing test is the first scientific attempt to test the abilities of AI; that is, to determine whether not a machine demonstrates intelligence. The test essentially works as follows: the test administrator asks a series of questions to two different interlocutors, A and B. However, A is a human being, while B is a machine. Their responses to the questions are returned as text to the test administrator. According to Turing, if a skilled administrator cannot determine which interlocutor is a human and which is a machine (based on the appropriateness of their responses), then the machine is exhibiting artificial intelligence. Now, given that the test administrator can ask anything of the machine (she could, in theory, just type gibberish), it seems that the standards for passing this test are sufficiently high. After all, it is unlikely that even a modern computer, which can execute billions of instructions per second, would be able to consistently come up with appropriate responses to the queries in a timely fashion. Additionally, the program required to handle all of this language processing would have to be incredibly complex. All things considered, the Turing test sets a high standard for artificial intelligence.

However, the Turing test is inherently flawed in such a way that it cannot truly measure the intelligence of a machine. The Turing test is essentially a behavioral test; that is, it measures the degree to which a machine succeeds in behaving like a human being. As such, it sets itself up for relatively straightforward counterexamples. One such counterexample, borrowed from Ned Block[1], runs as follows: suppose that we build a machine that is essentially an incredibly complex jukebox. For any given input, the program on this machine searches its database for the appropriate output. Now, if we could build a database of billions of input statements, each with their appropriate outputs, it is likely that this machine would pass the Turing test. However, this machine operates on the same principle as a jukebox, a machine to which we would be loath to ascribe intelligence. Thus, the Turing test, stringent as it is, it not a sufficient test by which to judge whether machine is exhibiting intelligence. The simple fact that a machine can behave like a human in certain, limited circumstances, is not sufficient to show that it is actually intelligent.

A more sweeping criticism of AI in general comes from John Searle, in the form of his famous 'Chinese Room' argument. The essential idea is this: imagine yourself (assuming that you do not understand any Chinese) in a room that has only the necessary items to create a rule-based input/output system. These items are: several baskets of funny shaped symbols (Chinese characters, unbeknownst to you), a very complex rulebook, and input and output slots. The rulebook contains rules for governing which symbols to put in the output slot, given the appearance of certain symbols in the input slot, and your internal state (e.g. whether or not you are in a state of having already received a certain input). Now, imagine that the rulebook is thorough enough so that you always can pass out an appropriate Chinese answer to an input question. This machine (you) would clearly pass the Turing test but it would not understand Chinese. The notion of understanding is critical here because Searle wants to ultimately claim that the Chinese room program cannot exhibit intelligence because it has only syntax (e.g. formal rules), and it cannot ascribe any meaning to the symbols that it confronts. Like Block, Searle creates an example in which there is a machine that can pass the Turing test, but it clearly does not exhibit intelligence. However, Searle's argument cuts deeper because it seems to show that intelligence could not possibly be programmed. The essential conclusion of this argument is that a digital computer is simply a computational machine and, as such, it can only 'interpret' syntax, but not semantics. The human mind, by contrast, interprets and makes extensive use of semantic claims in addition to syntactic ones. The ability to attach meaning (semantics) to strings of data (syntax) is one of the key features of the mind. Thus, since a digital computer does not have access to semantics, it cannot be a mind, regardless of the complexity of the program that it runs. This, Searle claims, refutes the central claim of what he refers to as 'strong AI'; the view that an appropriate program, with the correct inputs and outputs, constitutes a mind, regardless of the sort of hardware on which it is run (e.g. whether it is a program run by a brain or a microprocessor).

The Chinese room example certainly does provide additional reason to believe that the Turing test is not an adequate measure of machine intelligence. As Searle's example shows, a machine (or person, in this sort of scenario) could *act* as if they understood Chinese and thus pass the Turing test when there is clearly no such understanding present. But this is ultimately a problem for the Turing test, and not a problem for the prospects of AI.

The Chinese Room and the Possibility of Artificial Intelligence

The question that Searle really wants to answer with the Chinese room argument is the following: "Is instantiating or implementing the right computer program with the right inputs and outputs, sufficient for, or constitutive of, thinking?"[2] While the Chinese room argument certainly shows that implementing the right program is not constitutive of thinking, it does not show that this implementation is insufficient for thinking. Thinking is not simply running the appropriate program but, running such a program creates sufficient artificial brain activity to give rise to thinking.

According to Searle, all mental phenomena are caused by processes and states within the brain. Searle states this more explicitly as follows: "Mental phenomena, all mental phenomena whether conscious or unconscious, visual or auditory, pains, tickles, itches, thoughts, indeed, all of our mental life, are caused by processes going on in the brain."[3] For the sake of simplicity, I will follow Searle and abbreviate this idea with the phrase 'brains cause minds.' Searle is committed to this idea because it helps him to offer a solution to the mind-body problem. This theory has the consequence that the mind is essentially a sort of internal appearance of brain functions (that is, the way that we are aware of processes in the brain). While we are not directly conscious of the workings of the brain (e.g. I am not aware of how a certain dendrite is acting), we are conscious of the output of these processes, and it is this consciousness that forms the mind. However, it is very idea that also makes plausible a claim about the possibility of AI.

Searle claims: "It is essential to our conception of a digital computer that its operations can be specified purely formally…"[4] Further, "a typical computer 'rule' will determine that when a machine is in a certain state and it has a certain symbol on its tape, then it will perform a certain operation…"[5] That is, given a certain input and a certain discrete internal state, the machine will act in a certain way (that is, produce a certain output). But can't the operations of the brain be specified purely formally as well? The brain takes a certain input (say, particular sensory data) and, given a certain internal state, produces some output, which we are conscious of in the form of an appearance or thought. The operation of the brain is thus strongly analogous to the operation of a digital computer. But, Searle claims, "Minds are semantical in the sense that they have more than a formal structure, they have a content."[6] Thus, bearing in mind Searle's commitment to the idea that 'brains cause minds,' it follows that he would have to accept one of the following claims: "Since brains cause minds, and the brain has no semantic content, the mind can't have any either," or "Since brains cause minds, and the brain essentially is a digital computer, an appropriately complex digital computer can cause a mind as well." Notice that if the computer, coupled with its program *caused* a mind (as opposed to simply constituting one) this mind could have semantic content in the way that our minds do. While the artificial brain is defined wholly by its syntax, the artificial mind that it causes is able to add meaning (or at least give this appearance of it) to this formal structure, just like the human mind does.

Consider the first two premises that Searle employs in deriving his overall conclusion about AI. First, there is the claim that 'brains cause minds.' Second, Searle posits as a conceptual truth the claim that 'syntax is not sufficient for semantics.' Considered together, these two premises give rise to the following question: what is the source of semantics? It is relatively clear that the mind has access to semantics, but it is not so clear that the brain does. The brain can be described as having syntax; as having discrete states of molecules and electrons interacting so as to produce a mental process. These interactions are governed by formal rules (namely, the rules of biochemistry and physics). Thus, this syntax makes the brain a sort of formal system. But could the brain, a lump of wet, grey, biological matter, actually have a semantics as well? In other words, how could those discrete states of molecular and electronic interaction have a *meaning* in addition to their formal system? If Searle's argument is to make sense, it would also have to assume that there is some sort of meaning lodged in the particles of brain, which seems rather difficult to believe. It seems more plausible that the brain, like any other organ, is simply a syntactic system. Since, as Searle claims, syntax is insufficient for semantics, and the brain just is a syntactic system, it seems to follow that the mind could not have a grasp of semantics as, according to Searle, it is completely caused by the workings of the brain. But the mind clearly does have access to semantics and meanings, so at least one of the premises considered here must be false.

I believe that the false claim is the idea that syntax is insufficient for semantics. The universe is, in a sense, a rule-based system governed by syntax (e.g. the laws of physics), but semantics can be created within it (in the case of human minds). In most cases, it is true that syntax is not sufficient for semantics, as one would be hard pressed to defend the claim that a stone, for example, can have an understanding of meaning simply because it is governed by syntactic rules. However, I believe the solution here is that certain types of syntax, running on appropriate hardware, are sufficient for semantics. The complexity of the program running in the brain is sufficient to produce a conscious mind.

Further, Searle's argument seems to hinge on the fact that there is only one sort of mind. Indeed, he seems to believe that a human mind and a machine mind (if it were possible) would have to be the same sort of mind. Searle summarizes the view that he criticizes, strong AI as he terms it, by stating: "According to the most extreme version of this view, the brain is just a digital computer and the mind is just a computer program."[7] However, Searle claims later that proponents of strong AI believe that it's only a matter of time before technology develops artificial brains and minds and that "These will be artificial brains and minds which are in every way the equivalent of human brains and minds."[8] In the first case, Searle claims that strong AI holds that an artificial mind and a human mind would be *identical*, because they would be running the same program (albeit on different hardware). However, in the second case, strong AI merely seems to claim that an artificial mind and a human mind would be *equivalent* (presumably in the sense that they both exhibited the ability to think). In his argument, Searle only concerns himself with refuting the first claim, which is a very bold claim that is not easily defensible. By contrast, the second claim, that a human mind and an artificial mind can be equivalent, but not identical, is much more plausible. However, by refuting the first claim, Searle also seems to believe that he has refuted the second claim (perhaps because he does not acknowledge a difference between them). This confusion between the sorts of minds in which AI might be manifested helps Searle produce his misleading conclusion concerning the Chinese room.

The Chinese room argument is misleading because it asks us to compare the workings of an artificial brain with the workings of a human mind. The Chinese room is a completely formal, rule-bound system, as is the brain, whereas the workings of the human mind are not rule-bound. Given this unfair comparison, it is clear why it is so intuitive to claim that the Chinese room is not an example of intelligence. Brains, after all, are not intelligent, and the Chinese room is not even sufficiently complex to be a functioning brain. In short, the Chinese room does not represent a mind, which is, of course, what Searle intended to show with his example. However, this does not show that a digital computer cannot think. It only shows that a digital computer whose rules are not suitably complex to produce a mind cannot think.

Thus, it is consistent to agree with parts of Searle's conclusion and still hold that AI is possible. Searle's fourth conclusion is especially interesting as it even hints at this possibility: "For any artifact that we might build which had mental states equivalent to human mental states, the implementation of a computer program would not by itself be sufficient. Rather, the artifact would have to have powers equivalent to the powers of the human brain."[9] I take it that Searle means equivalent here in the sense that it functions to produce consciousness. So, the artifact would simply have to function to produce consciousness. The human brain does this, presumably, through a complex system of electrons and molecules interacting with one another. The artifact would do this via a complex system of electrons interacting with digital logic switches. But this artifact could still be made out of anything (including 'old beer cans powered by windmills,' to borrow Searle's derisive example). Searle is right to say that the mind is not a computer program, and that a computer program cannot be a mind. However, this is not the real issue. At the heart of the notion of AI is the idea that the mind is *caused* by a computer program, of sorts, in the brain and thus, it is clear that a computer program can *cause* a mind. This mind would not have to be like the human mind necessarily (indeed, how can we even know that there is a single 'type' of human mind that this new mind could be like?). In short, the mind is not simply complex hardware running a suitably complex program, but the brain is. Thus, if all mental activity is simply the internal appearance of brain activity, an artificial brain could produce a mind in much the same way that a real brain does.

Notes

[1] Block, 268-305.

[2] Searle, 36.

[3] Searle, 18.

[4] Searle, 30.

[5] Searle, 30-31.

[6] Searle, 31.

[7] Searle, 28.

[8] Searle, 29.

[9] Searle, 41.

Bibliography

Block, Ned. *Readings in Philosophy of Psychology*. Cambridge, Mass.: Harvard University Press, 1980.

Searle, John. *Minds, Brains and Science*. Cambridge, Mass.: Harvard University Press, 1984.