Indiana Undergraduate Journal of Cognitive Science 3 (2008) 51-57

Searle's Chinese Room Argument and its Replies: A Constructive Re-Warming and the Future of Artificial Intelligence

Mariano de Dompablo Cordio

Philosophy
University of Wisconsin, La Crosse

Abstract

This paper treats the philosophy of John R. Searle in his article "Minds, Brains, and Programs". It shows, using Searle's Chinese room argument (CR), that what Searle calls strong artificial intelligence (AI), the thesis that minds are to brains as computer software is to computer hardware, is not only false, but also that it must be false. The CR does this by arguing in effect that there can be no translation between Chinese and John's English understanding, and likewise neither can computers understand any question put to them because any question addressed to them is like the Chinese to John. On the other hand, ask John in English, which he understands as well as any native speaker, 'Do you understand Chinese?' and he will answer 'No.' What is the difference between John and computers? While John does not understand Chinese and does understand English, computers understand nothing. And because any programming language to computers is like Chinese to John, human-like behavior of a computer charged with running a robot, for example, proves nothing in the way of human understanding on the part of computers.

Because the CR does such a good job of proving the falsity of strong AI, a fundamentally different approach to the creation of AI is necessary. But, this is problematic for strong AI, namely, it leaves strong AI definitively behind.

1. Introduction

When we ask, 'Is artificial intelligence (AI) possible?' we really ask 'Can we create consciousness in computers?' This, as I see it, has been established by the philosophy of AI discussion so far. So, can we? And, if we can, what would we be

doing with the computer (if manipulating its inner structure or otherwise) in order for consciousness to be created with it? These questions put us in a place to define the AI project in two goals, a primary and a secondary. The secondary goal is a step toward the primary goal, consciousness in a computer. The secondary goal can be either of two possible alternatives as defined by the AI philosophy discussion so far: 1. writing a program, which when implemented in a hardware, results in a conscious computer according to test by observation OR 2. manipulating the inner structure of a computer in order to in effect reproduce within the computer the context for consciousness in a brain, providing the context necessary for consciousness but doing so in a computer. Both alternatives have an important background as they try for the primary goal. 1. takes consciousness (and a materialistic, biological account) for granted to a significant extent, and says that if behavior of the disputed conscious entity (the computer) can be mistaken for a human being given the proper circumstances, that entity understands (Turing 212). 1. says, 'Write the right program (one that programs for human-like behavior), and you have a conscious computer.' 2. assumes that consciousness is a property of the context provided by brains (and allows for a materialistic, biological account of consciousness).

In order to prevent any fudging on the definition of what strong AI (the thesis that minds are to brains as computer software is to computer hardware) supporters have as the goal of their projects, what I defined above as the primary goal, it is important to understand that strong AI projects and strong AI itself seem to be rooted in large part in science fiction. This happens in two ways. First, in so far as strong AI seems to borrow ideas for its projects like that of consciousness in computers from science fiction, it is innocuous. Ideas from science fiction examples of AI are not all that is borrowed. Second and more important is the very idea, the computer model of the mind, that hardware performing computation according to software-indicated directions could be intelligence, which may also be a holdover of too much science fiction, is damaging to a more biological account. But a deeper investigation into the origin of models of the mind over the course of history will have to be left for another time. To better pin down the goal of the projects of computer science from another direction, refer to the technical language used to name and describe computer scientists' AI projects. Some examples are: genetic programming, memory, master/slave programs, teaching programs, and programs that learn. All of these examples exude something of a human sense, a personification of their projects. In these examples, I have assumed that we see what computer scientists believe their projects can or will be able to do. But there is a problem with their usage of that language.

What computer scientists are trying to do with their projects is evident,

realize the primary goal, and that is illustrated by their language. The problem seems to be that computer scientists do not understand the significance of the language that they use, which is telling of an incomplete picture of the mental properties involved in their language e.g. consciousness or qualitative, subjective states of awareness and intentional states. So, to make a proper attempt at realizing the mind or its property consciousness in their computer would be almost happenstance because they do not understand the goals implicit to their language when, for example, they describe programs as teaching or learning. This taking for granted of one of these goals, consciousness, in the attempt to create AI reaches yet farther into the interdisciplinary field, cognitive science, of which computer science is a part. The philosopher branch of cognitive science as well often leaves out or inadequately treats in its philosophy of mind the What is it like? or the qualitative experience had by consciousness (Nagel 321). Conscious understanding and subjective intentional states are not, however, left out by the philosophy of John R. Searle in "Minds, brains, and programs" (235). His wonderful article is where this paper will begin.

2. Chinese Room Argument

Before continuing to my adapted rendering of the Chinese room argument appearing in Searle's article, the reader should understand that the Chinese room that Searle describes in his argument is designed to be identical in principle to any computer. Thus, anything that the Chinese room can or cannot do parallels all relevant computer capacities. A person, John, is in a room. John does not understand Chinese symbols nor is he capable of recognizing Chinese symbols such that he can distinguish Chinese symbols from Japanese symbols nor is there anything "To [keep John from believing that] Chinese writing is just so many meaningless squiggles" (Searle 236). Also in the room are two windows and two boxes. In Box 1 are directions written in English and divided into sets. Each particular set of English directions correspond to an attached Chinese symbol also inside Box 1. In Window 1 comes a Chinese symbol which John receives. John proceeds to Box 1, matches the Chinese symbol received to the same type Chinese symbol to which are attached a particular set of English directions. John follows these directions, which he understands as well "as any other native speaker of English", by finding in Box 2 the direction-indicated symbol (Searle 236). John then proceeds to Window 2 and puts this symbol out the window so that John effectively

> correlates one set of formal symbols with another set of formal symbols, and all that 'formal' means here is that [John] can identify the symbols entirely by their shapes . . . [answering] by manipulating uninterpreted

> formal symbols. (Searle 237)

And, all of this constitutes John simply behaving as a computer, performing "computational operations on formally specified elements" (Searle 237). So, John successfully performs computation without understanding Chinese. After all, his responses to this point have been indicated by the English directions. But what if we gave John questions written in English?

Suppose that we ask in English 'John, do you understand Chinese?' and indicate that he should pass his answer out Window 2. John would answer 'No.' What is the difference between John's response to the question in English and his response to the same question instead addressed in Chinese? Ask yourself what it would be like to be John, comparing both experiences; John experiences *understanding* of the English question and his answer. He understands nothing of the same question addressed in Chinese. Nor does he understand his Chinese answer as anything more than the pushing of a shape out of a window that reads "Window 2" over it. What does Searle provide us in this thought experiment? A distinction between human understanding of a familiar language versus a language one does not speak. It is also important to notice that computers and humans share an engagement with the shape of written words, but unlike John who understands the English and not the Chinese, computers cannot understand any language even though they work in them. These considerations provide us with an appropriate setting for a few of the many replies to the Chinese room argument.

3. Objections to the Chinese room argument

The first of the objections to the Chinese room argument to be treated will be the systems reply as named by Searle. The systems reply concedes that the man inside the Chinese room does not understand the Chinese version of the question put to him 'Do you understand Chinese?' However, John is but a part of the system, and the whole system, the Chinese room and any digital computer by way of its having at its disposal everything accessible to a digital computer, understands the question.

Searle replies, let John memorize every system element, the English directions and attached Chinese symbols in Box 1 and the Chinese symbols in Box 2 so that the composite's aspects comprise all aspects of the entire system. John still does not understand Chinese. There simply can be no translation between John's English understanding and the Chinese symbols, no matter John's Turing test mistaken understanding. And, "a fortiori neither does the system [understand], because" anything in the system is part of John (Searle 240). So, strong AI is false, the systems reply fails, and the Turing test has counterexamples, so it is

ineffective. With regard to the systems reply, Searle makes some additional remarks, which I think outline what has been the state of these affairs.

> It is not easy for me to imagine how someone who was not in the grip of an ideology would find the idea at all plausible. Still, I think many people who are committed to the ideology of strong AI will in the end be inclined to say something very much like this; so let us pursue it a bit further. (240)

The systems reply replies: "'the man as a formal symbol manipulation system' *really does understand Chinese*." (Searle 240) In this reply, the systems reply begs the question, that is, it insists the truth of its claims without argumentation in addition to its original argument. So, the systems reply is false. There are additional comments made by Searle with regard to objections to the Chinese room that do a great job of outlining the central errors of strong AI supporters.

The robot reply is the second objection to the Chinese room argument. It asks us to think about a new program. This computer, with its program written not only for the taking in and putting out of symbols, would perform the function of operator of the robot in which it is placed. The computer would operate the robot in such a way that its behaviors are similar and can be confused for something with human-level understanding. The idea is that this computer instead of the original digital computer treated by the Chinese room argument would have understanding. What should be noticed about the robot reply?

The robot reply indirectly claims that cognition is about being-in-the-world with certain realities of what it is to be a causal force in the world instead of just formal symbol manipulation (Searle 243). The obvious reply to the robot reply is that putting another computer inside of a robot does not get rid of the original problems outlined by the Chinese room argument; innovation in programming whatever cannot improve upon the problems which have thus far been outlined and are essential to all algorithm based approaches to the creation of AI, strong AI. In order to illustrate this criticism, imagine that inside that robot is John instead of the extra computer with its new program. The presence of John in the room for the purpose of carrying out the computation needed for the robot's operation is in principle equivalent to the needed computation otherwise being carried out by the computer. After having replaced the computer with John, follow through with the original Chinese room argument, that is, to the conclusion that there can be no translation between John's English understanding and Chinese, and one will understand that the robot reply is false. The implicit notion underlying the robot reply (as in the systems reply) that 'If it behaves like it, it must be it

(understanding)' does nothing to improve upon the essential state of John in the room and strong AI by the same force of the Chinese room argument. Given strong AI's falsity, an adjusted approach to the creation of AI will be necessary since strong AI must be false.

4. Further inquiry into the possibility of AI

A fundamentally different approach to the task of creating AI may be that of manipulating the orientation of the firings in the hardware of a computer in order to reproduce the necessary physical context for consciousness in brains but doing so in a computer. Though this approach may not reproduce the physical context necessary for consciousness because it would only reproduce the electrical portion of what would presumably be an electro-chemical reality in the brain. However, the electrical approach seems plausible in our pursuit of the reproduction of conscious subjective states necessary in order to in some way produce the understanding necessary for what is the precise primary goal of AI, intelligence. But this approach is problematic for strong AI, namely, it leaves strong AI definitively behind. Strong AI takes it as its assumption that hardware can realize certain desired properties of the brain without a neurophysiological account of the brain to reach the same goal. This adjusted approach leaves behind the use of just computer hardware in order to provide a physical context for consciousness, which is not just computer hardware proper. Without consciousness, however, intentional states would be had by nothing. So, while AI seems plausible with the appropriate context for consciousness, neurophysiology is required before it can be created. Strong AI cannot be true.

5. Conclusion

The original assumption of strong AI is that certain desired mental properties can be achieved using hardware in a computer instead of a brain, that brains are effectively hardware. But, this view, strong AI, has been proven false on multiple occasions by John R. Searle's Chinese room argument. The Chinese room argument has shown that there can be no translation between John's English understanding and Chinese, making understanding impossible at either John's level, that of the central processing unit (CPU), or at the level of the entire system, that of the CPU, the wires, and any additional hardware used in computation. Likewise, understanding is impossible when placing a computer inside of a robot so that it may move around in the world, for if John were in the room performing the same computational tasks, he could not understand Chinese, and moving around in the world in a human-like fashion does not improve upon the state of computers. The falsity of this last example, the robot reply, and the notion that mistaken behavior is a sure indicator of understanding casts serious doubt on the

Turing test. So, an adjusted approach is necessary for the successful creation of AI. The approach described in this paper, however, is a fundamentally different approach to that of creating strong AI, as manipulating the orientation of the firings in the hardware of a computer leaves the original assumption of strong AI behind, thus leaving strong AI definitively behind. Indeed, the re-orientation of firings in hardware on the model of a neurophysiologically defined orientation of neuron firings in order to provide the appropriate context for consciousness while using a materialistic, biological account of the brain is not hardware proper as required by the strong AI thesis. Rather, it is a synthesis of hardware and neurophysiology. This approach may not even reproduce the necessary physical context for consciousness because it only reproduces the electrical portion of what would presumably be an electro-chemical reality in the brain. So, AI (and not strong AI) is possible in principle, but it is dependent on a materialistic, biological account of the physical context for property consciousness as defined by a consciousness as a property of the brain hypothesis of the brain-consciousness relation. AI on the basis of the non-biological account of the brain in the strong AI thesis cannot be possible.

References

Nagel, Thomas. "What Is It Like to Be a Bat?" Philosophy of Mind: A Guide and Anthology. Ed. John Heil. Stamford, CT: Wadsworth, 2000.

Searle, John R. "Minds, Brains, and Programs." Philosophy of Mind: A Guide and Anthology. Ed. John Heil. New York: Oxford University Press, 2004.

Turing, Alan M. "Computing Machinery and Intelligence." Philosophy of Mind: A Guide and Anthology. Ed. John Heil. New York: Oxford University Press, 2004.