Francisca Vasconcelos

Minds & Machines

Recitation 1

3/24/2017

**The Chinese Room Argument and Systems Reply**

Despite major technological and algorithmic advances, one key area of dispute has remained unresolved throughout the history of AI: determining whether artificial intelligence can constitute true intelligence. Throughout this paper, I will 1) describe Searle's "Chinese Room Argument" on this topic, 2) describe the "Systems Reply" to Searle's argument, 3) describe Searle's response to the "Systems Reply," 4) lay a foundation in fundamental neuroscience in order to compare the human brain and Chinese room, 5) provide a case study of humans driving cars to explore Searle's notion of intelligence, and 6) use the human brain analogy and driving case study to support the "Systems Reply," disproving Searle's orignal "Chinese Room Argument" and "Systems Reply" response.

We begin by analyzing Searle's view on intelligence through his famous "Chinese Room Argument." As defined by Searle, Strong AI is "the view that all there is to having a mind is having a program," Weak AI is "the view that brain processes (and mental processes) can be simulated computationally," and Cognitivism is "the view that the brain is a digital computer." [1, pg. 22] Across several papers, he uses varying tactics, such as the "Primal Story" and "Casual Powers," to advocate a view opposed to all three of these. However, his strongest argument is that of the "Chinese Room" (which incorporates the idea of "Syntax vs Semantics"). This thought experiment consists of locking an English speaker in a room with a Chinese rule book. As slips of paper containing Chinese symbols are fed into the room, the English speaker writes a Chinese response, solely by following the instructions of the book. From an outside perspective, it appears that the mechanism inside the room has a total understanding of Chinese, giving coherent and appropriate responses as would a native speaker. However, the English speaker has no understanding whatsoever of the meanings of the words being passed in or written. In other words, the English speaker has an ability

to apply grammar rules to form coherent words, sentences, and conversations (known as syntax). However, he has no understanding of the actual meaning of the words, sentences, and conversations themselves (known as semantics). Without semantics, the English speaker has no intentionality. Searle further claims that this lack of understanding extends to the room as a whole. Because the Chinese room is analogous to a computer implementing a program, Searle argues that no computer could be capable of semantic understanding or intentionality. This argument can be formulated into a concise proof:

**Premise 1.** *The English speaker has a purely syntactical understanding of Chinese.*

**Premise 2.** *The Chinese room understands no more or less than the English speaker understands.*

**Conclusion 1.** *The Chinese room has a purely syntactical understanding of Chinese.*

**Premise 3.** *The Chinese room constitutes an appropriately programmed computer.*

**Conclusion 2.** *An appropriately programmed computer has a purely syntactical understanding of Chinese.*

**Premise 4.** *Understanding of Chinese is analogous to understanding of any phenomena.*

**Conclusion 3.** *An appropriately programmed computer is not sufficient for semantic understanding.*

**Premise 5.** *Mental content corresponds to semantic understanding.*

**Conclusion 4.** *An appropriately programmed computer does not have mental content.*

**Premise 6.** *Minds have mental content.*

**Conclusion 5.** *An appropriately programmed computer is not a mind.*

While the "Chinese Room" is a pinnacle argument of the anti-AI view, the "Systems Reply" is arguably one of the strongest counter-arguments. It asserts that "understanding is not being ascribed to the mere individual; rather it is being ascribed to this whole system of which he is a part." [2, pg. 419] The "Systems Reply" directly attacks **Premise 2** of the proof, with a negative and positive component. The negative component exposes Searle's flawed logic. It argues that if the English speaker is part of the Chinese room and the English speaker does not understand Chinese, it does not necessarily follow that the Chinese room

also does not understand English. For example, "we cannot reason from 'Bill has never sold uranium to North Korea' to 'Bill's company has never sold uranium to North Korea.' " [3, pg. 418] The positive component of the "Systems Reply" expands the Chinese room analogy:

$$Chinese\ room : computer :: English\ speaker : CPU$$

and thus proposes that "the whole system – man + program + board + paper + input and output doors – does understand Chinese, even though the man who is acting as the CPU does not." [3, pg.418] Although the CPU can be considered one of the primary computational components of a computer, a computer will not function without RAM, power, a graphics card, etc.

However, Searle's proclaimed "simple" response to the "Systems Reply" is far from satisfactory, prompting further discussion as to whether the Chinese room actually is "unintelligent." Searle's response argues that if the English speaker were to internalize the Chinese room, inheriting all its components (such as the rule book, slips of paper, writing utensil, etc.), he would have no better understanding of Chinese "because there isn't anything in the system that isn't in [the English speaker]." [2, pg. 419] That is, since the English speaker manipulates the objects in the Chinese room (which would otherwise do nothing on their own) to read input and create output, he is the driving force of the intellectual processes of the Chinese room. According to Searle, the English speaker's understanding is the only determinant of the whole Chinese room's understanding. It thus follows that, like the English speaker, the Chinese Room must have a purely "syntactic" understanding. However, if the Chinese room appears to have a full understanding of Chinese to an outside observer, does that not constitute an actual understanding of Chinese by the system as a whole? Why is it assumed that the other components of the Chinese room add no intelligence to the system, just because they cannot act without the English speaker? In fact, how does the way in which the Chinese room functions differ from the functioning of the human mind itself?

To draw a comparison between the Chinese room and human brain, we must first discuss biological neural processes. The brain is composed of individual neurons that fire in a specific sequence to carry out intellectual processes. Although the full workings of the brain are not yet understood, it is fair to make

the claim that each of these individual neurons do not have an understanding of the full process they are involved in. For example, a single neuron in the occipital lobe, involved in processing the sight of an apple, will not understand that the person has seen an apple or that an apple is red but can sometimes be green. Rather, it may find the edges of the apple and then work with all the other neurons in the entire visual system to create an understanding that the image is of an apple. Furthermore, the visual system may then interact with other systems, such as memory, to attribute ideas (i.e. apples are a type of fruit) to the visual input. The brain's ability to connect mental thoughts and processes is commonly referred to as human "intelligence."

An analogy between the Chinese room and human brain will now be used to prove that the Chinese room can have a "semantic" understanding even if the English speaker does not, undermining both the "Chinese Room Argument" and Searle's response to the "Systems Reply." In this comparison, it is fair to regard the English speaker as a single neuron or system within the brain. Other components of the Chinese room correspond to other brain systems. Now, consider the task of repeating a sequence of words. Since most animals cannot do this, it can be seen as a sign of "intelligence." Continuing the analogy with the Chinese room, incoming slips of paper correspond to the auditory system, the English speaker corresponds to the speech recognition system (which translates sounds into words), and outgoing slips correspond to the vocal system. The mental "rule book" translates incoming auditory neuron firing sequences into firing sequences necessary for the vocal system to produce words. In this way, the "rule book" determines the neuron-firing sequences of the speech recognition system. Note that the speech recognition system (English speaker) does not itself understand the "semantics" of language. In fact, it does not know anything about language. It solely uses a "rule book" to translate incoming neuron spikes into outgoing ones. Yet, as long as the brain (Chinese room system) is able to receive sound messages and produce the corresponding speech, most would consider it "intelligent." This is exactly the argument of the "Systems Reply" in stating that even if the English speaker does not have semantic understanding, the Chinese room as a whole can still have semantic understanding. If one considers human speaking capabilities "intelligent," one must

also consider the Chinese room's ability to respond like a native speaker "intelligent." It thus follows that a "semantic" understanding must be inherent to the Chinese room system as a whole, even if it is not inherent to the English speaker. This directly contradicts Searle's claims that the Chinese room is not intelligent (made in the "Chinese Room Argument") and that the English speaker encapsulates the full understanding of the Chinese room system (made in his response to the "Systems Reply")

To further argue Searle's response to the "Systems Reply" we must fully understand his notion of intelligence, which will be achieved through a case study of driving. In the previous argument, it was shown that many "intelligent" human behaviors are not considered "intelligent" in Searle's view because they do not actually involve "semantic" understanding. Rather, the system as a whole (part of which is the human) must be attributed the "intelligence." Consider, now, the task of driving. Most would find human ability to operate a piece of machinery as sophisticated as a modern car, a sign of human "intelligence." Yet, for this to be valid by Searle's standards, the human driver would have to understand the "semantics" of vehicle locomotion. Most people, however, have no understanding of how a car works or how torque applied to a steering wheel results in a change of direction of the car. Therefore, by Searle's logic, they only have a "syntactic" understanding of how to drive. Through experience (learning how to drive) people simply internalize a function ("rule book") that maps steering wheel torques to wheel orientations. Most people cannot even explain what this function is or how they are able to implement it. All they know is that when they perform certain actions they get certain responses from the car, allowing them to drive. This is as far from a "semantic" understanding as one can get. In fact, the engineer who builds the car and has full knowledge of all relevant "semantics," is not necessarily a better driver than a professional race car driver, who has a limited understanding of the "semantics" but learned a better "rule book" by trial and error. Hence, under Searle's view, the fact that a human can drive does not make him or her "intelligent."

Drawing upon this car example, either Searle's views expressed in his original "Chinese Room Argument" or in his response to the "Systems Reply" can be proven false. In the previous paragraph, a human's "syntactic" understanding of driving was found to be "unintelligent" in Searle's view. However, since most

people refer to human ability to drive as an "intelligent" ability, this poses a direct contradiction between Searle's meaning of intelligence and the general meaning. If this is the case, can we not regard a "syntactic" understanding as a full understanding (equivalent to "semantic" understanding)? If so, we have disproven his "Chinese Room Argument" claim that the Chinese room is not actually intelligent. If not, let us now consider the "system" composed of the car and human inside. This system converts what the human sees (the input) into complex combinations of thrust and torque by the car (the output). This conversion is precise enough to allow the system to meander through tortuous mountain roads, drive around the occasional soccer ball that bounces into a parking lot, etc. As argued by the "Systems Reply," although the driver only has a "syntactic" understanding of how to drive, "semantic" understanding is built into the car itself, allowing the system as a whole to act in an "intelligent" manner according to Searle's view. However, contrary to the view expressed in Searle's "Systems Reply" response, "semantic" understanding in each of the individual components of the system (in this case, the human driver) is not necessary for the system as a whole (the car and human driver system) to have full understanding. Therefore, if we are willing to challenge Searle's view of syntactic vs semantic understanding, we can discredit his original "Chinese Room Argument." Otherwise, we can use the car-human system to discredit his response to the "Systems Reply."

In order to uncover flaws in Searle's notion and implementation of intelligence in his "Chinese Room Argument," the function of neurons in the brain and human-car relationship in driving were analyzed. Following the "Systems Reply," it was shown that not all components of a system must be intelligent in order for the whole system to be intelligent. These pro-AI views are reinforced by recent technological advances, such as deep- and reinforcement-learning systems, which have proven that we are capable of creating intelligent digital systems by mimicking the functionality of the brain. Who is to say, though, that our intelligence is the only form of intelligence? As in the words of Alan Turing, "may not machines carry out something which ought to be described as thinking but which is very different from what man does?" [4]

## References


**1.** J. Searle. Is the Brain a Digital Computer? *Proceedings and Addresses of the American Philosophical Association*, 64(3):21–37, November 1990.

**2.** J. Searle. Minds, brains, and programs. *The Behavioral and Brain Sciences*, 3:417–457, 1980.

**3.** N. Block. The Mind as the Software of the Brain. *An Invitation to Cognitive Science*, 1995.

**4.** A. M. Turing. Computing Machinery and Intelligence. *Mind A Quarterly Review of Psychology and Philosophy*, 59(236):433–460, October 1950.