The Problems with the Chinese Room

Nicola De Angeli, nicola.deangeli@mail.polimi.it

August 2020

Abstract

The Chinese room argument advanced by Searle has been influencing the strong AI literature for decades. We define strong AI and general AI, distinguishing the two concepts. We briefly describe Searle's ideas, focusing on the Chinese room thought experiment. We analyze and build upon some of the critiques advanced by other researchers in the strong AI community. We find the Chinese room argument to be flawed when confronted with the concept of simulation and the ideas of functionalism, suggesting that there is still no good reason to believe that a computer cannot attain consciousness.

1. Introduction

Searle in [7] argues that the biological matter of the human brain is essential for intentionality. The centerpiece of Searle's argument is a thought experiment, the Chinese room, where a person that follows the rules of a program answers to questions in Chinese without having intentionality of any of the Chinese utterances. The Chinese room has now been influencing the strong AI literature for decades. In this paper, we first define strong AI, distinguishing it from general AI, as the two terms are often confused with one another. Then, after a brief description of Searle's argument, we analyze and build upon some of the critiques advanced by Stevan Harnad in [2] and Georges Rey in [6] to challenge Searle's ideas. We find the Chinese room argument to be flawed when confronted with the concept of simulation and the ideas of functionalism; consequently, there is still no good reason to believe that a computer cannot attain consciousness. We attempt to thoroughly understand Searle's thought experiment and convincingly highlight its problems to hopefully help shift the focus of future strong AI research onto other (arguably more interesting) issues.

2. General AI versus Strong AI

When discussing the topic of strong AI, one needs to first define and distinguish the possible theoretical forms of machine intelligence clearly. It is often the case that the concept of strong AI is confused with that of general AI (AGI). The literature, however, shows that these two concepts are substantially different.

In [1], Ben Goertzel claims that, while the precise definition of general AI is vastly disputed within the AGI community, there is a broad agreement on some key ideas that characterize the concept. Therefore, general "involves the ability to achieve a variety of goals, and carry out a variety of tasks, in a variety of different contexts and environments". Moreover, "the general intelligent system should be able to handle problems and situations quite different from those anticipated by its creators" [1, p. 2]. Such a system can accomplish the above by generalizing the knowledge gained through experience, "to transfer this knowledge from one problem or context to others" [1, p. 3]. The concept of general AI is usually contra posed to that of narrow AI. Analogously, Kurzweil defines narrow AI in [4] as the creation of systems that carry out specific "intelligent" behaviors in specific contexts. Consequently, by changing ever so slightly the context or behavior specification of the system, some human reprogramming is needed for the system to retain its level of intelligence.

To provide an example, imagine a robot that is in charge of taking care of plants in a garden. The creators of such a robot might provide it with some prior information on the plants it will take care of, such as when to water them, or which compost to feed them. Let us now suppose that one day this robot is charged with taking care of a different, never seen before kind of plant that is not known by the creators. If the robot is endowed with general intelligence, it will be able to adapt to the new type of plant, perhaps by experimenting different

treatments for the first few days, but also considering previous general knowledge such as "when it is very sunny outside, the plant needs more water". Moreover, the robot will gain additional knowledge when taking care of the new plants, which will be useful in the future. In contrast, a robot provided with a narrow type of intelligence, if not reprogrammed, would treat the new plants the same as the old ones, with detrimental effects on their health, and will gain no additional knowledge by doing so.

Of course, given the example above, it is evident that there exist several levels of general intelligence. Indeed, human intelligence can be considered more general than the intelligence of the robot in the example. Still, the AGI community believes that the general intelligence displayed by man is not the maximum expression of general intelligence [1].

Strong AI is a term that has an established meaning in the AI and cognitive science literature. Searle defines a strong AI as an intelligent artificial machine that is able "to understand and have other cognitive states" [7, p. 417]. Consequently, we can define a weak AI as an artificial machine that, although capable of displaying intelligent behavior, is not able to understand and have other cognitive states. Note that Searle uses the term "cognitive states" to refer to the states of the mind that allow for intentionality, that is, the ability of humans and animals to be conscious, be about things, represent things, and give meaning to representations.

General and strong AI are thus two very different concepts. Indeed, the human brain is an example of a system that is endowed with both strong and general intelligence. In the machine learning literature, the branch of meta learning attempts to develop a general intelligence by training neural models to easily adapt to multiple tasks [3]; on the other hand, strong AI is not actively pursued. Finally, one could be tempted to say that a strong intelligence must also be at least as general as the intelligence displayed by humans or animals. However, we argue that this is not the case. In principle, there is nothing particularly absurd about the existence of a living being with a very narrow intelligence and ability to be intentional about things. However, we do concede that such a living being would probably need to be created artificially, as its inability to adapt to new conditions and tasks would be unfavorable in the context of natural selection. Therefore, we conclude that strong AI and general AI do not necessarily imply one another. In the following sections, we only consider strong AI, which is the subject of Searle's critique.

3. Searle's Argument

Searle's argument in [7] is directed against the predominant philosophical positions of the strong AI community at the time, that is, the conviction that "the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states". He refutes the above idea by claiming that: "[i]nstantiating a computer program is never by itself a sufficient condition of intentionality" [7, p. 417]. The centerpiece of Searle's argument is the Chinese room thought experiment.

- Inside the Chinese room:
    1. there is an English-speaking person, a book with a set of rules written in English, some empty sheets of papers, and two breaches on the walls, one for input and one for output;
    2. the person inside receives some sheets of paper with meaningless symbols from the input breach;
    3. the person can correlate one set of formal symbols with another set of formal symbols;
    4. the person applies the rules to fill other sheets of paper and provides them to the output breach.
- Outside the Chinese room:
    1. stories and questions about the story in Chinese written on sheets of paper are entered in the room from the input breach;
    2. after some time, the appropriate answers in Chinese exit the room from the output breach.

Searle claims that, despite ultimately providing a correct answer as output, the person inside the room does not understand anything of the stories written in Chinese; in other words, the person does not have *intentionality* of the content of the stories. He then observes that the person, by mindlessly following a program that processes Chinese scribbles, can be considered as analogous to a modern computer following the instructions given by a program. Because of this, he finally concludes that a computer cannot possibly cause intentionality, as the Chinese room argument shows that formal symbol manipulation – which is all the computer does – is not enough to produce cognitive states. Instead, one needs to consider the hardware running such programs before determining its ability to think. One cannot separate the program from the brain's hardware and call it a mind, because the brain's hardware contains something that is also crucial for the mind to exist.

The Chinese room argument has been the subject of numerous critiques, with some having gone as far as to define the field of cognitive science as the ongoing mission of demonstrating Searle's arguments wrong [5]. We describe, analyze, and build on some critiques that we believe are effective in questioning the validity of the above claims in sections 4 and 5.

4. On Simulations

Harnad presents multiple arguments to challenge the Chinese room. One of them is the "Simulation versus Implementation" argument, where he argues that "Searle fails to distinguish between the simulation of a mechanism, which is only the formal testing of a theory, and the implementation of a mechanism, which does duplicate causal powers" [2, p. 12].

Searle seems to acknowledge to an extent the importance of programs and simulations in cognitive science, stating that he has "no objection to the claims of [the] weak AI [community]", as in "the principal value of the computer in the study of the mind [being] that it gives us a very powerful tool" [7, p. 417]. However, unlike Searle, Harnad provides a detailed analysis of the term "simulation".

Harnad contraposes the concept of simulation to that of implementation. He imagines a computer simulation of flight, taking into consideration all known aerodynamic factors, where models of planes can be tested. He then supposes that, after finding a plane that can fly in the simulation, the same plane is reproduced in the real world and is shown to also be able to fly. From this example, we can infer that simulation is abstract, theoretical, and formal, while implementation is concrete, practical, and physical. Nonetheless, "both contain the relevant theoretical information, the relevant causal principles" [2, p. 2].

The idea of simulation is then applied to analyze Searle's Chinese room. Inside the room, a person simulates *something* and produces behavior that is the same as that of a person understanding Chinese. That *something* the person is simulating is, according to Harnad, not understanding Chinese, but rather *the simulation of understanding Chinese*. In summary, "Searle's "simulation" only simulates simulation rather than implementation" [2, p. 12]. Finally, because "[t]he simulation of understanding Chinese does not understand Chinese any more than the simulation of flying flies" [2, p. 4], it is only natural for the person inside the room to not understand a word of Chinese. In conclusion, Harnad's argument recognizes Searle's Chinese room as a simulation and proves the inadequacy of the thought experiment.

While we believe in the validity of Harnad's conclusions, we believe the argument to present many issues. First, we argue that Harnad confuses the simulation with the simulation of the simulation when explaining why the person does not understand Chinese. In theory, one could only claim that *the simulation of the simulation of understanding Chinese does not simulate understanding Chinese*. Still, we argue that there is no reason to believe that such "nested simulation" would result in understanding Chinese either.

Furthermore, we do not believe that what the person is simulating is a simulation of understanding Chinese. A simulation would need to be informative about the implementation; however, Searle does not provide any detail on the type of program the person is simulating in the room[1] other than that the obtained behavior is the same as the behavior of a Chinese speaker; the use of a mere (astronomically large) look-up table could replicate such behavior without providing any insight on how the Chinese speaker understands.

[1]We refer to the original Chinese room formulation

Finally, even if the program inside the room were to be the program of understanding Chinese[2] (supposing such program exists), we argue that the person inside the room would still not understand Chinese: they would not *implement* the program, but rather *simulate* the program, and the simulation, as we have seen before, does not understand. We thus consider the Chinese room to be flawed either way. We believe that the problem resides in the person, which is only capable of simulating the program provided in the room.

The three following questions naturally arise from our considerations.

1. What is the person inside Searle's Chinese room actually simulating?
2. What does it mean for a person to *implement* a program?
3. Why can the person inside the room only *simulate* the program provided in the room?

We address these points by drawing parallels with the work of Georges Rey [6] in section 5.

5. On Functionalism

Georges Rey makes an important distinction between functionalist and behavioral theories of the mind. While behavioral theories only appeal to the input and output of the system, functionalist theories also consider the system's internal state. Rey claims that strong AI is a functionalist theory: a system that is not mediated by the right sort of program "will not be regarded as satisfying some mental predicate, no matter how much its behavior may resemble the behavior of a system that does" [6, pp. 2-3].

Consequently, we can affirm that a program of behaving as if understanding Chinese is not necessarily also a program of understanding Chinese. On the other hand, the program of understanding Chinese is surely also a program of behaving as if understanding Chinese. We will refer to such programs as the *Chinese behavior program* (CBP) and *Chinese understanding program* (CUP).

We believe that, in the original Searle's Chinese room, what the person inside the room is simulating is the CBP: it contains the information to act as if understanding Chinese, but not necessarily the information to really understand Chinese. One could argue that this is a contradiction: the simulation of behaving in a certain way should not behave in a certain way no more than the simulation of flying should fly, and yet the Chinese room behaves correctly. However, we argue that in the case of the behavior of a program as in its input/output correlation, it is indeed simple, starting from the simulation of the program, to behave like the program: we just need to consider the inputs and outputs of the system as the inputs and outputs of the simulated program, which is exactly what the person inside the room is doing[3].

Even when the program inside the room is the CUP, we argue that understanding of Chinese is not achieved by the person. The problem is that "the rules are still outside the person in the room: he has to look up the rules in a book" [6, p. 5]. For the person in the room to understand Chinese, they would have to run the CUP directly on their brain. Using Harnad's terminology, the person is *simulating* the program rather than *implementing* the program. A person implements the CUP when that program runs directly on their brain[4].

However, we believe that a human has no way to directly run a program written on paper on their brain. When learning Chinese, we do not study the very program we aim to run in our brain but rather read from books that embellish raw rules with examples and connect new information with prior knowledge, which has been empirically proven to be effective in generating the execution of the correct program on the brain. If we were to be shown just the program, we would not be able to run it directly on our hardware as a computer does, because there is no reason for us to be able to do so[5].

Consequently, we can confidently claim that the person inside the room is incapable of doing anything other than simulating the program, which causes the person to be unable to understand, even in the case in which

[2] We define the program of X as the program that, when run by a machine, causes X.
[3] This concludes the answer to question 1.
[4] This concludes the answer to question 2.
[5] This concludes the answer to question 3.

the program inside the room really is the CUP. However, this does not negate the fact that running the CUP on hardware does cause Chinese understanding, both in the case of the brain and – arguably – the computer.

Our claims until now assume the existence of the CUP, that is, the program that, when running on the brain, causes the understanding of Chinese. However, we argue that such a program may not exist. In fact, according to Rey, "[t]o put Searle's example even in the running as a possible counterexample to Strong AI, we need to imagine the person in the room following rules that relate Chinese characters not only to one another but also to the inputs and outputs of (...) other programs (...) to account for the other mental processes of a normal Chinese speaker" [6, pp. 3-4]. If we consider the brain to be a collection of running programs, then the program(s)[6] in charge of the Chinese language may need to rely on other ones to cause the understanding of Chinese. As Rey explains, we "no more need[] or ought to ascribe any understanding of Chinese to this latter part of the entire system than we need or ought to ascribe the properties of the entire British Empire to Queen Victoria" [6, p.6]. The Chinese room thus also fails in addressing the possibly multi-program nature of the brain: even if a machine inside the room were to run the exact same program that is also run to handle language on the brain of a Chinese speaker, that machine would still not run the possible other programs necessary for intentionality.

6. On the Internalized Chinese Room

Various *prima facie* counterarguments are addressed and rebutted by Searle himself in [7]. The most common rejoinder, the so-called "system reply", proposes the idea that it would be not the person inside the room themselves, but the system, to understand Chinese. In response to such an argument, Searle formulates another version of the thought experiment where the person in the room internalizes all the elements of the system. Because the person still does not understand anything of the Chinese stories, and the person is the system, Searle claims to prove the fallacy of the system reply. The conclusion is, according to Searle, not surprising, as it would be quite extraordinary for simple, inanimate elements of the Chinese room to be crucial for the existence of a mind. Though one cannot disprove that a priori, there is simply not enough evidence to believe that is the case.

Indeed, considering the ideas we have expressed so far, Searle's reply seems to be a menacing counterargument. If the person were to internalize the CUP, such a program would be stored in their brain. If the person inside the room were then to start running the program, we would conclude that the program is running directly on the brain, which should imply understanding: a contradiction.

However, we argue that there are two plausible ways to internalize the CUP inside the brain. In the first case, the program is *run directly on* the brain and communicates with the other programs run on the brain fittingly. In this case, we believe that the person can indeed understand Chinese. Consequently, we argue that Searle's reply focuses on the second case, where the program is instead *memorized on* the brain as a sequence of instructions. In this case, the brain does not *run* the program, but rather, again, *simulates* memorized operations by running a program that is equivalent to the program a modern computer runs to operate a virtual machine: while the behavior obtained might be the same on the surface, a simulation of the program, and not the actual program, is running on the hardware. Because simulations of understanding cannot understand, we can conclude that it is reasonable for the person to not understand Chinese. The contradiction mentioned above is thus only apparent.

7. Conclusions

We provided the definitions of several types of machine intelligence, distinguishing between general and strong AI and showing that each one does not imply the other. We then described Searle's Chinese room argument against strong AI and provided counterarguments to it based on the works of Harnad in [2] and Rey in [6]. Our belief is that the program provided in the room is not the same as the one necessary for understanding, and even if it were, the person inside the room would still not be able to understand, as that program would not run directly on their brain. Furthermore, we argue that Searle fails to account for the possibly multi-program nature of the brain. We

[6] Multiple programs may be needed to handle language as well.

addressed the internalized version of the Chinese room, resolving an apparent contradiction in our claims. Finally, we proved that the Chinese room does not question in the slightest the validity of strong AI as a functionalist theory of the mind. There is still no evidence that modern computers cannot attain consciousness.

8. Acknowledgements

I would like to thank professor Viola Schiaffonati for her revision and helpful suggestions. I would also like to thank Alberto Archetti for his time and valuable discussions of the ideas presented in this paper.

References

[1] Ben Goertzel. Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1):1–48, 2014.

[2] Stevan Harnad. Minds, machines and searle. *Journal of Experimental & Theoretical Artificial Intelligence*, 1(1):5–25, 1989.

[3] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*, 2020.

[4] Ray Kurzweil. *The singularity is near: When humans transcend biology*. Penguin, 2005.

[5] Margery M Lucas and Patrick J Hayes. *Proceedings of the Cognitive Curriculum Conference*. University of Rochester, 1982.

[6] Georges Rey. What's really going on in searle's "chinese room". *Philosophical Studies*, 50(2):169–185, 1986.

[7] John R Searle et al. Minds, brains, and programs. *Behavioral and Brain Sciences*, pages 417–457, 1980.