A critique to Searle's Chinese room argument

Luca Molteni
Student of Computer science and engineering
Philosophical issues of computer science course
Politecnico di Milano
Student code: 913275
Email: luca7.molteni@mail.polimi.it

Searle's notorious Chinese room argument, has been for a long time considered as the definitive proof against the attempts of Strong AI to make computers and programs think. This work led to a split between those who actually claim that the argument is sufficient proof to put an end to the aspirations of the functionalists and those who criticize it. What follows is a critical analysis of Searle's work aimed at highlighting some critical points in his reasoning. The first criticism is directed at the vague and ambiguous definition reserved for the main topic of the discussion: intentionality. Such semantic weakness gives raise to a number of unanswered questions and inflames the debate even more. Subsequently, it is highlighted what appears to be a logical flow in Searle's reasoning that would seemingly lead to a conclusion in opposition with his claims. As a matter of fact, it can be shown that the human simulation of a computer described in the Chinese room gedankenexperiment actually understand something; in fact, as Searle itself says, the system understands the program, which is a set of rules written in English, in the same way a human understands. As a third point the very nature of his argumentation and the repercussions derived from starting from different assumptions with respect to his are considered; It is shown that by embracing a vision in accordance with epistemological solipsism his results would be invalidated. Finally, the conclusion takes into account a reflection about the current state of AI and what has changed since Searle's original publication.

1 Introduction

The purpose of this paper is to critique Searle's arguments presented in the work *Brains, minds and computers*. [1] Searle goes after the principles of the so called Strong AI by arguing that: 1) Computer programs are never by themselves a sufficient condition for intentionality and 2) Intentionality in human beings is a product of causal features of the brain and computer does not have such causal features. Strong AI, on the other hand, claims that the computer, if appropriately programmed, is really a mind, in the sense that it can be said to understand and have cognitive states similarly to a human brain.

Searle argues against this claim by means of a notorious Gedankenexperiment called the Chinese room argument that will be later explained in detail since it is of absolute interest for this critique. It is of utmost importance to have clearly in mind that in this dichotomy between Searle and Strong AI I do not want to root for either part; I will rather try to show how Searle's argument could be not as effective against the possibility that computers have cognitive states as it seems to be.

2 The chinese room argument

First of all the Chinese room argument is a Gedankenexperiment, that is, a fictual mind experiment carried out with the support of imagination. [2] With this kind of reasoning Searle tries to show how a human simulating the behaviour of a computer running a program is not understanding anything and that consequently no computer does as well.

I will try to describe it as faithfully as possible to the original presented in the 1980 paper. The experiment goes as follows. A man who does not understand a single word of Chinese but understands English is locked in a room. For him Chinese symbols are just scribbles, in the sense that he would not even be able to

tell whether they are Chinese or Japanese symbols instead. He receives 2 batch of Chinese writings and a set of rules in English. The man is also given a third batch of Chinese symbols and some more instructions in English allowing him to correlate elements of the first two batches with the third one. These rules tell him how to produce some Chinese symbols in response to the content of the third batch.

All the previous elements can be reinterpreted in the following manner; quoting Searle:

> *Unknown to me, the people who are giving me all of these symbols call the first batch 'a script', they call the second batch a 'story' and they call the third batch 'questions'. Furthermore, they call the symbols I give them back in response to the third batch 'answers to the questions ' and the set of rules in English that they gave me, they call 'the program'.* (Searle, 1980, p. 3)

The reasoning proceeds explaining that such system would be able to answer questions in Chinese in the same way a Chinese speaker would do and indeed if some men were to observe only the input and outputs of the system without knowing what is inside they would reasonably conclude that the system is able to understand Chinese. The point is that for the man inside the room "it seems quite obvious that he does not understand a word of the Chinese stories since is just performing manipulation of formal symbols" as Searle states. (Searle, 1980, p. 3)

Now that the basis of this argumentation have been set up I can proceed to show how these apparently clear and logical conclusions are the result of premises and reasoning that can be considered at least questionable.

3 The importance of understanding intentionality

The first problem, which makes it more difficult to debate about Searle's argument, regards the definition of understanding and intentionality, since the ambiguity of these concepts leaves plenty of room for interpretations. About the latter Searle simply says: "Intentionality in human beings (and animals) is a product of causal features of the brain. I assume this is an empirical fact about the actual causal relations between mental processes and brains. It says simply that certain brain processes are sufficient for intentionality". (Searle, 1980, p. 1) It leaves no hints whatsoever about what is intended with causal features of the brain or with mental processes/states and it does not explain why and how human and animal brains are so special to have such features. Searle's notion of intentionality is briefly described as: "that feature of certain mental states by which they are directed at or about objects and states or affairs in the world". (Searle, 1980, p. 15) About the definition of understanding it is said that it implies a combination of possession of intentional states and the truth of these states. He further specifies that for the purpose of discussion only the possession of the states is considered thus making possession of intentionality or understanding equivalent. (Searle, 1980, p. 14) Since this paper is based on a critique of Searle's original work I will remain faithful to this position and the words "intentionality" and "understanding" will be adopted without implying any substantial difference.

The poor description of the phenomena in which the substance of Searle's argument ultimately lies gave raise to a number of critiques from the philosophical world. As the Stanford encyclopedia highlights, where the term "intentionality" is concerned, there is always confusing and contentious usage. [3] Margaret Boden in a section of her 1988 book, *Computer Models of the Mind*, goes as far as saying that intentionality is not well-understood, reason to not put too much weight on arguments that turn on intentionality. [4] Howard Gardiner holds that Searle owes us a more precise account of intentionality than he has given so far, and until then it is an open question whether AI can produce it, or whether it is beyond its scope. Until the discussion is based on these fuzzy bases the dispute between Searle and his critics is not scientific, but almost religious. [5] Paul and Patricia Churchland agree with Searle that the Chinese Room does not understand Chinese, but hold that the argument itself exploits ignorance of cognitive and semantic phenomena. They raise a parallel case of "The Luminous Room" where someone waves a magnet and argues that the absence of resulting visible light shows that Maxwell's electromagnetic theory is false. [6] As it has been highlighted, without a clear description of the concepts of understanding and intentionality, Searle's argument has become more ambiguous and manipulable.

I do not pretend to be able to provide a definitive or even sufficiently acceptable definition of what intentionality is, but the fact that Searle has dismissed the most important concept of his argument with so little effort leaves the doors open to manipulations and disputes. It could be exactly this conceptual ambiguity that has led the Chinese room argument to be one of the most debated topic in the world not only of Philosophy of artificial intelligence but also Philosophy of sciences in general, with as many as 4.910 publication regarding

the topic in the period between 2015 and today available via Google Scholar. And perhaps, precisely for this reason, I am now writing about this subject.

4 About the understanding in Searle's human computer

Searle claims that it is reasonable for the man inside the room that he is not understanding Chinese; this being the proof that a computer and a computer program are not by themself a sufficient condition for understanding. (Searle, 1980, pag. 3) In my opinion, even if we accept the fact that the man's feeling of him understanding nothing is true, there is a fallacy in Searle's reasoning. He clearly states that in his argument the man is simulating the behaviour of a computer. Quoting: "I simply behave like a computer, ...I am simply an instantiation of the computer program." or "in the Chinese case the computer is me". He also claims that the man in the room actually understand the English rules he is given, as he mentions in the description of the experiment: "The rules are in English, and I understand these rules as well as any other native speaker of English". (Searle, 1980, pag. 3)

Here comes the problem, Searle could have even proved that the man in the room, the computer, does not understand Chinese, but to do so he claimed that the computer understands the program, which are the English rules. This would mean that inside the computer there is indeed understanding and that computers have causal or mental states consequently disproving his argument. One could argue that the man implementing the program has of course mental states, but that we should just refer to the portion of him necessary to set up the experiment, such as the ability to manipulate formal symbols. Still, in order to do so, there is need of understanding of the English rules, since they are the program itself. If we were to take away the assumption that the computer understands English, the experiment could not be carried out in the same way, since if it was not for the understanding of the rules the man would not have been able to correlate the Chinese symbols. That means that the understanding of English is part of Searle's human implementation of a computer as much as the ability to manipulate formal symbols. Note that I have been alternating the use of the word man and computer, since in the context of the Chinese room experiment, according to Searle, the man itself IS the computer. Searle's goes on saying: "we can see that the computer and its program do not provide sufficient conditions of understanding since the computer and the program are functioning, and there is no understanding." (Searle, 1980, p. 4) The problem is, as I mentioned, that there is indeed understanding, since as Searle itself said, the program is understood by the computer.

Now, there are two possible paths to follow; if we accept Searle's implementation of the computer as the human in the room described in the Chinese Room argument we should then conclude that in reality computers and programs have understanding, giving away some points to Strong AI supporters. On the other hand we could simply conclude that maybe Searle's human representation of a computer fails in its first objective, that is the one of simulating a computer, thus proving nothing about the incorrectness of Strong AI suppositions.

5 About the claims of understanding

What if the assumption that Searle makes about the man in the room not understanding anything is wrong? After all the whole Chinese room situation is a mental experiment; How can Searle be so sure that in reality, if one were able to memorize and understand the rules to correlate the Chinese symbols, he would still have the perception of not understanding?

Perception itself is a problem. Searle says that the man in the room feels that he does not understand anything but in such way Searle is simply highlighting the human awareness of intentionality. Daniel Dennett suggests that Searle conflates intentionality with awareness of intentionality. Searle has apparently confused a claim about the underivability of semantics from syntax with a claim about the underivability of the consciousness of semantics from syntax. [7] Others have noticed that Searle's discussion has shown a shift from issues of intentionality and understanding to issues of consciousness and that Searle links intentionality to awareness of intentionality. [8]

To support his argumentation Searle exploits multiple time the presupposition that men are capable of predicting whether there is understanding or not, for example relying on the knowledge of the inner workings of things. I will proceed showing that this position has a strong impact in his argumentation and that if were to embrace different assumptions Searle's conclusions would be very different. The new point of view that I want to follow is generally defined as Solipsism. [9] Solipsism is a philosophical idea claiming that only one's own mind is sure to exist leading to all knowledge residing outside one's own mind being unsure.

The external world and the other minds cannot be really known but are just a projection of our own perceptions. [10] Many different varieties of solipsism have developed through time. Metaphysical solipsism, for example, goes as far as saying that the world and other minds do not exist; [11] this particular vision will not be taken into consideration since it wont lead to interesting results regarding Searle 's argument. In fact such extreme positions put severe limitations to what can be really known since subjective impressions become the sole possible and proper starting point for philosophical construction. According to the more moderate Epistemological solipsism only the directly accessible mental contents of the solipsistic philosopher can be known. [12] This concept clashes violently with Searle's supposition that men are able to "see" intentionality in other beings. Now, assuming that any man has the sufficient introspective capabilities to realize whether he understands or not, that is whether his mental states can be intentional or not, according to solipsism this ability should be confined only to the ego itself.

It can be seen that Searle premises and conclusions are now put on the line, starting from the presupposition that intentionality in humans and animals is a product of the causal features of the brain. In fact it is not certain anymore whether other humans and animals have intentionality, let alone it being a product of causal features. The intentionality that a man attributes to other people is simply the result of the projection of what we see in ourselves onto anyone else who shows similar features and behaviors but the reality is that we really do not know anything except about ourselves. As Dennett states all intentionality is derived. Attributions of intentionality, to animals, other people, even ourselves, are instrumental and allow us to predict behavior, but they are not descriptions of intrinsic properties. [13] Getting back to the Chinese room argument, it now appears absurd thinking of being able to attribute or not intentionality to the system on the basis of the perception that the man in the room has about himself. As Kurzweil says we can't know the subjective experience of another entity [14], and in the case of the Chinese room the man can only understand and draw conclusions about his human nature, not the one of a computer. It is easy to see in the Chinese room what we want to be seen; strong AI supporters that would like to assert the possibility of creating understanding using a programmed digital computer could simply perform the same gedankenexperiment and conclude that the man in the room feels that he understands. The implication seems to be that minds generally are more abstract than the systems that realize them, such as in the example of the man in the room, and it is not possible to address the issue of intentionality just by knowing its inner workings or worse just by its external appearance.

If the reality is that we cannot know anything about the external world and the intentionality we see in other humans and animals is the result of the transposition of our own intentionality to entities that appear similar in both aspect and behaviour, why we should not apply this same rules to other artifacts such as the computer? The system in the Chinese room is able to make external viewers think that he is understanding Chinese. Why, still, we are doubtful about its possession of understanding then? Probably there is a strong bias that prevents us from conceding intentionality to something that appears to be too different from us. Searle itself, while answering the so called combination reply agrees on the fact that a robot whose appearance and behaviour was indistinguishable from a human would be certainly granted with the attribute of intentionality. He then says that as soon as we knew that the behavior was the result of a formal program, and that the actual causal properties of the physical substance were irrelevant we would abandon the assumption of intentionality. (Searle, 1980, p. 9) Nevertheless under the light of the solipsistic view these arguments do not hold anymore since the presence of causal features and the physical substance of things became irrelevant for the attribution of intentionality itself.

6 Conclusions

I hope that the last three paragraphs have been able to explain how and why Searle's Chinese room argument is such a debated topic. It ultimately exploits a generalized ignorance and ambiguity about the concepts of intentionality and understanding; It presents some logical flaws since his computer simulation has been shown to have indeed some understanding. Finally, its argument is based on the strong assumption that intentionality is only derived from the causal features of the physical brain; assumption, that as it has been indicated, appears to be excessive and not reasonably justified. Starting from different premises it has been made clear that Searle's conclusions are no longer so definitive.

To better showcase the skepticism and uncertainty that still surrounds Searle's claims there is a multibillion project co-founded by the European Union called the *Human Brain Project* . It is expected that in 2023 it will be able to provide a complete simula-

tion of the brain functioning that will be used to understand neuropsychiatric diseases such as depression and schizophrenia and to build controllers for truly intelligent robots with broad applications in industry, services and the home. [15]

To conclude: has the CR argument effectively given such a punch in the face of Strong AI as someone says? If we look at the present it doesn't seem so. Many things changed since Searle publication (1980). Computational power has increased exponentially. What was once considered the best hardware now fits an area smaller than a coin. Things that were not even imaginable at the time of the Chinese room dissertation are now available for our everyday wonder. AI moved from the so called winter to a new era where tangible and useful results are achieved every day. Ultimately none of the two part can declare their claim as true, but it seems that lately strong AI has some reasons to say that it may be right.

References

[1] Searle, J. R., 1980. "Minds, brains and programs". *Behavioral and Brain Sciences,* **3**(3), pp. 417–57.

[2] Perkowitz, S., 2010. Gedankenexperiment. `https://www.britannica.com/science/Gedankenexperiment`.

[3] Cole, D., 2019. "The chinese room argument". In *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, ed., spring 2019 ed. Metaphysics Research Lab, Stanford University.

[4] Boden, M. A., 1988. "Escaping from the chinese room". In *Computer Models of Mind*, J. Heil, ed. Cambridge University Press, pp. 238–251.

[5] Gardner, H., 1987. *The Mind's New Science: A History Of The Cognitive Revolution*. New York: Basic Books.

[6] Churchland, P. M., and Churchland, P. S., 1990. "Could a machine think?". *Scientific American,* **262**(1), pp. 32–37.

[7] Dennett, D. C., 1978. *Brainstorms: Philosophical Essays on Mind and Psychology*. Bradford Books, p. 336.

[8] Chalmers, D. J., 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press, pp. 115–165.

[9] Thornton, S. P. *Solipsism and the Problem of Other Minds*. University of Limerick. Internet Encyclopedia of Philosophy https://www.iep.utm.edu/solipsis/.

[10] Avramides, A., 2019. "Other minds". In *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, ed., summer 2019 ed. Metaphysics Research Lab, Stanford University.

[11] Angeles, P. A., ed., 1992. *Harper Collins Dictionary of Philosophy*. Harper Perennial, New York.

[12] Wikipedia contributors, 2019. Solipsism — Wikipedia, the free encyclopedia. [Online; accessed 24-May-2019].

[13] Dennett, D. C., 1987. "Fast thinking". In *The Intentional Stance*. MIT Press, pp. 324–337.

[14] Kurzweil, R., 2002. "Locked in his chinese room". In *Richards 2002*. pp. 128–171.

[15] Markram, H., Meier, K., Lippert, T., Grillner, S., Frackowiak, R., Dehaene, S., Knoll, A., Sompolinsky, H., Verstreken, K., Defelipe, J., Grant, S., Changeux, J.-P., and Saria, A., 2011. "Introducing the human brain project". *Procedia Computer Science,* **7**, 12, p. 3942.