# Searle's Chinese Room Reconsidered

By Niccolò Aimone Pisano

**Abstract**

In this essay, I will argue that Searle's (1980) argument in its original version does not debunk the computational theory of mind, but it does if adequately modified in the light of Dreyfus's (1992) argument. I will first outline the core ideas behind the view that is meant to be challenged by Searle's mental experiment; then I will describe the thought experiment itself, also exposing, and objecting to, Boden's (1987) reply which highlights two difficulties of the argument, namely SIR's (Searle-In-the-Room) understanding of English and Searle's biological chauvinism. Finally, I will show how taking into consideration Dreyfus's (1992) claim that the computational theory of mind entails some unavoidable regress about the rules to be applied allows Searle's argument to be a more effective attack to the computational theory of mind.

## Introduction

In this essay, I will argue that Searle's (1980) argument does not debunk the computational theory of mind, but it does if adequately modified. I will first outline the core ideas behind the view that is meant to be challenged by Searle's mental experiment. Then I will describe the thought experiment itself, also exposing Boden's (1987) reply which highlights two difficulties of the argument; namely, the alleged *petitio principii* due to Searle-In-the-Room's understanding of English, and the unwarranted reliance on the in principle impossibility for a machine to simulate human cognitive behaviours. In doing so, I will show how the former objection is really weaker than the latter. Finally, I will show how taking into consideration some ideas from Dreyfus (1992), concerning the practical impossibility to specify in advance the rules to be followed by a computational machine in order for it to pass the Turing test, allows Searle's argument to be a more effective attack to the computational theory of mind, in that in this way it hinges upon a proper argument without simply relying on the intuitions conveyed by his thought experiment.

## The computational theory of mind

First, it will be useful to spend a few words to explain what the computational theory of mind challenged by Searle's argument consists of. Roughly, a computational machine is a device which formally, i.e. purely syntactically, manipulates symbols following a set of rules and taking into account the state the machine is in. In other words, on the basis of the inputs received and depending on its internal conditions, it can give outputs according to the instructions previously implemented (the programme). For instance, if there is a rule such as "if you read *how are you?*, answer *fine* if you are in good conditions or *not so good* if you are not in good conditions", every time the machine will receive the input *how are you?*, it will give *fine* as an output if its internal state is regular, *not so good* otherwise. Now, the computational theory of mind consists in understanding the human mind as a computational machine. This means that every human behaviour can be interpreted as the output of the application of some conditional rule, where the inputs are a combination of external stimuli (e.g. the question *how are you?*) and pre-existing internal states (e.g. the actual health of the person the question is asked to).

The most important claim of the computational theory of mind addressed by Searle is that it is possible for a complex enough machine not only to simulate human cognitive performances, but to be considered endowed with the same cognitive skills as the human mind in virtue of this. Such a view can be articulated in two points:

*(1)* Human mind can be considered a computational machine.

*(2)* If behaviour B cannot be distinguished in type from a previously observed behaviour A, and we do not possess other information about B except for that gathered from its observation, we have to consider both as outputs of the same type of underlying processes.

I have already explained what *(1)* means. On the other hand, *(2)* can be clarified by an example: "(a) x has opened her umbrella" & "(b) x opens her umbrella whenever she notices that it is raining" → "(c) x has noticed that it is raining"; hence, *(2)* entails that whenever (a) occurs it is necessary to conclude that (c), unless instead of (b) we have, say, "(d) x told me that she wants to prove the falsity of the superstition according to which it is bad luck to open an umbrella when you are inside a building".

From these assumptions it follows that if one were capable to build a machine complex enough to simulate all human cognitive processes, so that it were behaviourally indistinguishable from a human in every situation, we would have to consider such machine as endowed with a mind similar to the human one, in that it would always produce the same sort of phenomenal outputs. This is what the Turing test for cognition is: if a human, after a conversation with a machine, is not able to tell whether his interlocutor is another human or a machine, the machine can be said to have passed the test and can be considered to be thinking like a human.

## Searle's thought experiment

Now that I have explained what position Searle's thought experiment aims to challenge, let us turn to it. Searle imagines that there is an English speaker (SIR, Searle-In-the-Room) inside a room, completely isolated except for the possibility to receive messages from the outside, to which he is required to write replies using an English rulebook. These messages, as well as the answers, consist of strings of symbols completely meaningless to SIR. It turns out that they are Chinese characters, and the rulebook allows SIR to answer meaningful Chinese questions as correctly as a native Chinese speaker would. Now, given *(1)* and *(2)*, the external Chinese questioners are induced to suppose that whoever is inside the room is understanding Chinese. But, Searle argues, this is not the case: SIR does not know that he is manipulating Chinese symbols, nor that he is having a conversation with someone: he is just following rules such as "if you see *squiggle*, write *squoggle*". Thus, the conclusion of the thought experiment is that, since SIR's processing symbols in a strictly syntactical way can be exhaustively equated to a machine's activity, although a machine can pass the Turing test (as SIR does), it does not have any proper understanding of what it is doing. Therefore, since they differ at least in that they can understand things, human minds cannot be instantiations of computational machines. In other words, the fact that in such a scenario the machine proves that the Turing test is not a valid test for cognition, in that it can be passed without there being any proper understanding of the dialogue, leads to the conclusion that *(2)* fails. But the failure of the second claim of the computational theory leaves *(1)* unwarranted; hence, the computational theory of mind has to be abandoned.

## Boden's objections

There are two difficulties with this thought experiment, as Boden (1987) has pointed out. First, it is not true that there is no understanding at all: SIR does genuinely understand the English rules written in the rulebook, so that, although his consequent behaviour may not be displaying any understanding, the equation "SIR = computer" is not correct.

Second, Searle claims that there are intuitively compelling reasons to hold that the mind has peculiar features in virtue of its biological realizers; therefore, no non-biological machine can possibly be endowed with equivalent cognitive skill, since different materials do not support them (Searle 1980, p. 421). In other words, Searle claims that even if his thought experiment failed, there are material reasons for rejecting the computational theory of mind, which unacceptably equates human minds and computers. However, Boden argues, there is no clear evidence that this is the case: (so far) we do not know why the brain allows the existence of human mind, thus not being available any explanation, not entirely based on intuition, of why other brain-like structures could not allow it as well. Therefore, holding that artificial brain-like machines could not support human minds seems not to be enough well-supported a claim to be acceptable.

Boden claims that the more important of these two objections is the former, since the latter concerns an easily dismissible claim. I disagree: what is more important is that the intuition behind the second of Searle's claims, if adequately backed up, is what allows to debunk the computational theory of mind. Indeed, it is not hard to modify Searle's thought experiment in order to avoid Boden's first objection, the "English reply": it will suffice to replace SIR with a system of levers and pulleys that purely mechanically elaborates the output. In this way, no human would be involved, and yet we would not grant to the mechanism inside the room a proper understanding of Chinese. In other words, the English reply is focused on SIR's humanity, while its sole purpose is to more easily convey the intuition on the basis of which the thought experiment is concocted. We could as well replace SIR with a mechanism whose ignorance of Chinese language we have ascertained before its implementation in the thought experiment. SIR's humanity only makes clearer that, even if inside the room there were something *per se* capable of proper understanding, that specific input-output dynamic would not be evidence for SIR's understanding of Chinese.

## Dreyfus and the infinite rulebook

Before replying to Boden's other objection, I will introduce some ideas elaborated by Dreyfus (1992), whose main argument against the computational theory of mind concerns some unavoidable regress about the rules to be applied. In what follows, I will make use of these ideas to reinterpret Searle's thought experiment to make it more effective an attack to the computational theory of mind.

In the first part of this essay I have defined a computational machine as a symbol-manipulating device which operates following syntactic rules that take into account the inputs and the current state of the machine. Simple as it may seem, this pattern is extremely difficult to be realized when it comes to machines able to pass the Turing test. For, as Dreyfus stresses, it is easily applicable only to the most recent form of human language, the formalised one, but it is practically impossible to specify a set of rules exhaustively covering the complexity of the use of natural languages. Since they extensively rely on context-dependent contributions, even a simple question such as "where is the cat?" (not to mention sentences involving indexical terms, that is, terms such as "this" or "here" which entirely depend on the context) would require exceedingly complex computations to be answered, in that it would be necessary to specify things such as what sort of object a cat is, which particular cat we are talking about, what kind of landmarks can be mentioned in the answer, and so on. Therefore, in the light of Dreyfus's argument, there are only two alternatives: either it is necessary to specify a thorough regulation of all the infinitely many linguistic combinations; or it is necessary to specify an infinite set of rules about when and how to apply a finite set of comparatively simple rules. For instance, on the basis of the first alternative, in the case of the question "where is the cat?", we might need a set of extremely specific rules such as "if *where is the cat*, and the person-shaped object knows that you own a cat-shaped object, and you are in a building with such and such features, …, then answer "it is in the living room" if the cat is in the living room, or "it is in the kitchen" if it is in the kitchen, …".

Both these tasks cannot obviously be materially accomplished: the first would require an infinite number of "atomic" rules, since the number of the sentences that can be uttered in any natural language is infinite; the other would require a potentially infinite generative process of rules' specification, for analogous reasons.

Now, Searle's argument does not take into consideration this fundamental issue at all. The English rulebook is taken to be comprehensively covering all the answers to the questions asked by the external Chinese speakers, and the focus of the thought experiment is rather on the intuitive unacceptability of labelling SIR's activity as a proper form of understanding Chinese. Nonetheless, an appeal to intuitive disagreement does not count as a satisfactory rebuttal of a theory. This is the reason why I do not think that Searle's "Room" argument debunks the computational theory of mind. However, if it is adequately modified in order to accommodate the aforementioned issue, it does.

## Searle's thought experiment modified

In what follows, I will present my alternative version of Searle's Chinese Room, which also counts as a reply to Boden's second objection, as it will emerge. Suppose that instead of a question-answer dynamic, Searle described a more general command-execution one. It could be ordered something like "the next time I say *x*, answer *y*". In this case, instead of applying a previously encoded rule such as, "whenever you read *x*, write *z*", it would be more appropriate to apply a rule such as:

if *the next time I say x, answer y*, write in the rulebook "if you read *x*, write *y*" and apply

"if you read *x*, do not apply 'if you read *x*, write *z*', apply 'if you read *x*, answer *y*' and 'if

you have applied 'whenever you read *x*, write *y*' delete that rule from the rulebook'".

It is clear that it is impossible to write such a rulebook, since, as I have shown before following Dreyfus, it would be needed either an infinite number of "atomic" rules, or an infinite set of rules about when and how to apply some simple rules. Therefore, in order to effectively attack the computational theory of mind, it is better to concede that the Turing test is a valid test for cognition (in accordance to what *(2)* prescribes), as the upholders of that theory claim, and instead argue that, since as a matter of fact human minds exist while no cognitively comparable computational machine could be built, nor even in principle, human minds are not computational machines (against *(1)*) In other words, it seems to be a more effective strategy to attack *(1)* while accepting *(2)*, instead of accepting *(1)* and challenging *(2)* as Searle does.

Moreover, suppose that a supporter of the computational theory of mind decides to face the second of the two aforementioned infinite regresses. He could argue that the rules-related regress can be stopped by assuming that there is some pre-normative (i.e. antecedent to and independent from the rules) bottom-level where the input-internal state couple physically determines the outputs (in this case, the formulation of the rules). That is, the regress mentioned by Dreyfus can be avoided by assuming that there is no need for either an exhaustive enumeration of all the rules or of all the rules for generating the rules: at some point, SIR's production of outputs would be determined by a purely physical, non-linguistic, causal process. But this is where Searle's point which is underestimated by Boden kicks in. If at this basic level the rules, understood as action-reaction patterns, are determined by the physical properties of the substances constituting the machine instead of by linguistically formulated rules, the possibility that cognitive performances might be carried out by a non-human system just as effectively as by a real human is ruled out in virtue of the different physico-chemical properties of distinct substances. It seems, then, that a supporter of the computational theory of mind can ground the rules-generating process only at the price of conceding that no computational machine can ultimately be able to have the same rules-generating process of human minds, because that process would strictly depend on the physical properties of the substances the brain is made of. But if minds are to be conceived of as a particular kind of computational machines, they cannot display properties that no machine can possess. However, if my argument is correct, they do. Therefore, *(1)*, the most important of the two central tenets of the computational theory of mind, has to be abandoned.

## Conclusion

It is finally possible to re-assemble all the pieces. Searle's original argument does not debunk the computational theory of mind, intended as the view whose core ideas are *(1)* and *(2)*, since it merely makes appeal to intuition in order to challenge *(2)* while accepting *(1)*. However, some modifications in the set-up of the thought experiment, together with a shift in its focal points justified by Dreyfus's (1992) argument, allow to more strongly attack the computational theory of mind by accepting *(2)* and rejecting *(1)*, while at the same time avoiding Boden's objections. In fact, on the one hand, the English reply can be easily avoided by replacing SIR with a mechanical system; on the other hand, since on the basis of Dreyfus's argument it is impossible to specify in advance a set of rules that would enable a machine to pass the Turing test, the fact that human minds can actually pass it entails that there must be some pre-normative process (i.e. some process antecedent to the rules) in act. If that is the case, then Boden's second point, according to which there is no guarantee that the physico-chemical properties of the brain are what ultimately leads to the occurrence of proper human cognitive behaviours, must be abandoned. But this means that no artificial computational machine can actually pass the Turing test, given that in order to do so it is necessary to have the same physico-chemical properties of human brains. Therefore, the computational theory of mind can be defeated by Searle's Chinese Room.

## REFERENCES

- Boden, M. A. (1987). *Escaping from the Chinese Room*. University of Sussex, School of Cognitive Sciences.
- Dreyfus, H. L. (1992). *What computers still can't do: a critique of artificial reason*. MIT press.
- Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417-424.