The Chinese Room and the Bête Machine. An epistemology of two thought experiments

Rispens, Sybe Izaak

Abstract. John Searle's famous thought experiment known as the „Chinese Room argument“ (CRA) is arguably the 20th century's greatest philosophical polarizer. Countless arguments supportive and critical of the CRA have been published in the last few decades. Yet what is missing so far, is an epistemological analysis on the role of the thought experiment in Searle's argument, compared with the 17th century bête machine thought experiment of René Descartes. This article proposes such analysis, arguing that the CRA is not fit to support sound philosophical debate because of its conflicting nature as a deconstructive thought experiment.

INTRODUCTION

At the height of the AI debate, in 1980, John Searle wrote a famous essay called "Minds, Brains and Programs". [1] The essay features a well-known thought experiment, which is now commonly referred to as the „Chinese Room argument“ (CRA). Countless arguments supportive or critical of Searle's argument have been published in the past few decades. Yet what is missing so far, is an epistemological analysis that draws historical parallels to Descartes bête machine thought experiment. [2, 3] This historical parallel allows for a meta-analysis on the role of the thought experiment in Searle's argument. Such analysis can direct future research questions in Cognitive Science.

After briefly summarizing the Chinese Room argument in part I, I will classify it in a three-level systematization of thought experiments (Part II). The classification will allow for drawing historical parallels to another famous thought experiment of the same category, formulated by René Descartes (Part III).

This historical meta-analysis of the CRA will lead to the conclusion that its chosen thought experiment was bound to fail. That will lead to three meta-guidelines that may guide future research within cognitive science.

I. The Chinese Room

In “Minds, Brains and Programs” it is Searle's aim to criticize a position he calls "strong AI". Strong AI claims that the computer is not merely a tool in the study of the mind, but rather, that the appropriately programmed computer which functions according to formal structures, really is a mind.

Searle starts out by acknowledging that a computer is able to "think", in the sense that it is able to process information by formal syntactical rules. Yet, the computer cannot be intentionally related to the objects and states of affairs in the world, even if these reflect the machine's current mental states.

After discussing several other thought experiments, Searle then introduces his own thought experiment: the „Chinese Room argument“.

Briefly summarized, the argument goes like this: a native English speaker is being locked into a room, after which he is given batches of Chinese writing. The native speaker does not understand Chinese writing, but a set of rules in his own language allows him to identify the Chinese symbols by their shape. Based on the formal rules he has been given, he is able to answer written questions in Chinese in a way that is indistinguishable from a native Chinese speaker.

As a result, the native speaker in the thought experiment passes the Turing Test. He is able to convince the native Chinese reader with his sensible answers, yet he himself does not understand anything of the symbols he wrote.

The morale of the thought experiment is this: take the native speaker in the Chinese Room as a metaphor for a computer that performs computational operations on formally specified elements. Then, it must become apparent that, like our native speaker, a computer does not “understand” the symbols given to it. The set of rules that it operates upon, does not enable the machine to understand them like a human being would. So, even if the computer would pass the Turing Test, there would not be any understanding by virtue of running a program of the right sort.

If Searle would have stopped here, his CRA would probably have entered the history of the AI Debate as one of many other arguments. Yet, Searle gave his readers much more than a thought experiment. He added something that we may now see as a skillfully written open invitation to the reader to think for him or herself. He continues his thought experiment with something best described as an aloud self-reflection by the author upon his previous thought experiment. What follows is an extensive set of possible answers, both critical and in favor of his own position. He takes the reader by the hand in a process of "Selbstdenken" - thinking for oneself. However, Searle's start into the process of Selbstdenken was such, that it triggered a polarizing intellectual debate. [4]

Unfortunately, this debate has generated more heat than light. Therefore, I refrain from any reflection upon the argument itself. Instead, I propose a meta-reflection on Searle's instruments for Selbstdenken: the thought experiment. The central question then becomes of a meta-question: what type of thought experiment did Searle introduce in the CRA, and was this type suitable for demonstrating the question at stake?

---

1 Institut für Wissenschafts- und Technologiekommunikation, Berlin, Germany, email: s.rispens@iwtk.eu
II. The CRA as a thought experiment

Early in the 1990s, James Robert Brown proposed a taxonomy for thought experiments in the natural sciences. [5]. The taxonomy consists of three classes, which are useful for analysing the nature of the CRA:

1. deconstructive thought experiments - show the logical contradictions in premisses. Typically, the destructive thought experiment is some sort of reductio ad absurdum of a pre-existing theory. Schrodinger's cat is for instance an example of a purely destructive thought experiment. It is meant to show that the Copenhagen interpretation of quantum mechanics is in flagrant violation of well-entrenched common sense. The fate of the cat's life in the sealed box depends on the state of a subatomic particle. According to the Copenhagen interpretation, the cat would remain both alive and dead to the universe outside the box, until the box is opened. So, this thought experiment illustrates the counterintuitiveness of quantum mechanics and the mathematics necessary to describe quantum states.

2. constructive thought experiments - support a theory by pointing out their necessary, yet not sufficient arguments. An example is Maxwell's thought experiment in support of the second law of thermodynamics. The aim of the thought experiment is to demonstrate that it is possible that differences in temperature in an isolated physical system equal out through heat passing from a cold body to a hot one - a thought held to be absurd at the time. In the thought experiment a demon is introduced, which controls a door between two adjacent chambers, one hot, the other cold. The average speed of the molecules is different in the two chambers; there are, however, some fast ones in the cold room. These the demon lets into the hot room while letting slow molecules from the hot room into the cold. In this way heat has passed from the cold to the hot chamber, since the average speed of the molecules has increased in the hot room and decreased in the cold. Therefore, it is indeed possible that heat can pass from a cold to a hot body.

3. platonic thought experiments - refute all theories but one. They are simultaneously constructive and destructive. They destroy the old and they create the new in a single blow. Leibniz's argument for vis viva or living force - a concept we now call kinetic energy - is an example. His thought experiment simultaneously destroyed the Cartesian view of what is conserved and at the same time, established a new principle of conservation of energy.

At first sight, the CRA seems to be a classical deconstructive thought experiment. It takes the idea of the Turing Test, and then brings in the common sense notion of what it means when a human being "understands a language". This common sense notion is then put in contrast with a human being passing the Turing Test based on computational operations on formally specified elements. One may interpret this as a reductio ad absurdum of the strong AI notion, that an appropriately programmed computer, which functions according to formal structures, really is a mind. Yet, there is more to it.

Unlike Schrödinger's purely destructive thought experiment of the cat in the sealed box, the CRA is not exclusively deconstructive. The description of the room and the rule-based operations of the native speaker are machine analogies. Such machine analogies, as will be outlined below, are always constructive, not deconstructive. To make this point, we will briefly draw a parallel to René Descartes' idea of a bête machine.

III. 17th century parallel: The Bête Machine

Descartes' thought experiment commonly known as the "bête machine" is one of the most famous ideas of machines of the 17th century. Like all thought experiments, it starts with a typical 'if', in the sense of 'suppose', 'assume', 'imagine'. What follows is a hypothetical machine setting that gives a fair amount of technological details. These details are important, as they guide what instinctive knowledge, or common sense, is bound to 'know' intuitively from experience about the issues at stake.

The central idea of the "bête machine" will be summarized here without all mechanical details and also without an analysis of the set of "common knowledge" that these details would relate to for the 17th century reader.

"If there were such machines..." part V of the Discourse on Method states, "...having the organs and the shape of a monkey or of some other animal that lacked reason, we would have no way of recognizing that they were not entirely of the same nature as these animals." [8]

This is the core of the bête machine thought experiment. It gives constructive evidence for one theory, namely that animals are complex clockworks. The central notion is that if one assumes that technology will advance in such a way that more parts and more complex mechanisms will be available over time, then that will lead to more delicate behavior, and in the end, that development will lead to such scale of complexity, that it will match the external behavior of a monkey.

There are three underlying assumptions here. One: technology will advance continually in the sense of the advancement of the Baconian Sciences. The mechanical clockwork represents the idea of "state-of-the-art technology": the most advanced technology of the time. Two: there is a positive correlation between technological advancement and external behavior. That means, in ten, twenty or fifty years, the technological proficiency will be such that indeed, a bête machine will be able to simulate the external behavior of a monkey. Three: there is no fundamental difference between biological and mechanical matter. This is a typical thought for early modern philosophy: the natural world can be seen as a gigantic mechanical clockwork whose laws of motion can be known with mathematical precision. Thus, the only distinction between a machine and a monkey is a graduate one in terms of functional complexity.

This is a perfect example of a machine-based thought experiment. It shows very well why these types of thought experiments must be constructive. The mechanical clockwork serves as a "hook" for common sense notions of matter, perception, behavior, complexity and technology. These are constructive intuitive connotations, and they all contribute to the constructive, intuitive evidence for an existing theory. It works through double negation: the imagined possibility of the bête machine is an argument for the exclusion of its impossibility.

If Descartes would have tried to prove that the bête machine would not be possible, his thought experiment would prove per definition that not the idea itself, but the technological setting of the thought experiment would have been wrong. The thought experiment in which the bête machine does not live up to the complex behavior of a monkey, is simply a bad thought experiment, because, for instance, the proposed machinery is outdated, too simplistic or naive. A deconstructive machine-based thought experiment would prove only one thing: that the chosen technological analogy is not complex enough. It would
not prove that the presuppositions of its underlying theory - the mechanical philosophy - in itself were contradictory or erroneous.

After Descartes presented his theory of bête machine and the according thought experiments, the issue was widely discussed in philosophical circles of the time, such as those of Arnauld and Mersenne [5]. Descartes himself entered a heated debate with Libert Froidmond, who questioned — next to various religious matters — dozens of mechanical details in Descartes' thought experiments. Froidmond’s basic intuition seems to have been that the Cartesian invisible gearwheels of nature, of which only size, position and movement count, would not be complex enough to explain for all animal behavior. Similar doubts had Alphonse Pollot, another one of Descartes’ philosophical correspondents, who wondered whether animal affections and passions could accurately be explained with the mechanism provided in Descartes’ thought experiments. 2

Descartes answered by presenting yet another thought experiment. More vivid and detailed than the first one, and more compelling. Imagine, he answered Pollot, a man who never saw any other living beings than other humans in his life. This man would be asked to build mechanical animals and creatures with a human-like figure. Then he would have two ways of distinguishing human-like automats from real humans: first, mechanical man would move much more perfectly and regularly than biological man. Indeed, Descartes was convinced that machines were far superior in making elegant and precise movements than humans. Second, mechanical man would not be able to answer sensibly to arbitrary questions, as would biological man. These are, as Descartes argues, our two sure criteria to distinguish machine from man. However, that ignorant observant would have no way to distinguish machine from animal. Both, mechanism and animal flesh would fly away when disturbed, yelp when kicked, flee when seeing a stick with which it had been beaten, and bark at own will, while ignoring all human questions.

The seventeenth century saw many such mechanical thought experiments, yet not even one was deconstructive. On the contrary, mechanical philosophers were so confident in their ability to think up a rational and mechanical explanation for all seemingly inexplicable or occult phenomena, that they were even encouraged to find mechanical explanations for phenomena that could not be directly explained with isomorphic structures known from the most advanced technology of the time. For instance, the historian of science can easily find multiple mechanical explanations for such phenomena as why Saturn sends his dark moods to earthling man through invisible mechanical reduction gears.

The key issue is, that if the CRA will point out any contradiction in theory, then these contradictions can per definition be explained away by challenging the analogous relationship between "machine" or "set of rules" (in this case, the Chinese Room as an analogy for a computer) and "theory". And this is exactly what happened in the CRA-reception: many authors have shown that the Chinese Room would be a too simplistic setting, or, as Margaret Boden has done, to show that interpreting formal rules should also be understood as 'understanding' [7]. It has been argued that the details for the computational operations have been too simplistic. So the issue is here, that any deconstructive thought can be addressed to issues related to the details of the "Chinese Room machinery", or to definitions of concepts used in the thought experiment. They do not touch the underlying theory.

So, if Searle wants to demonstrate that strong AI is not a good model, or does not explain the functionality of the human mind, or takes a too narrow sense of 'understanding' for granted, then that is due to limitations in the setting of the Chinese Room and is common sense notions.

In order to be convincing to his critics, Searle then continued on a path Descartes already took in the 17th century. He added increasingly more details to his thought experiment. And every new detail would refer to new common sense notions of human reasoning or notions of what it means to be an intelligent agent. For example: with every detail added to the thought experiment, more background notions of what it means to "understand" language, or how formal rules may lead to intelligent behavior, are added to the setting.

This results in two major epistemological flaws in the Chinese Room. First: it becomes circular reasoning. That what needs to be proven (what it means to "understand" something) becomes part of the Chinese Room and its associated common sense notions of it. Second: whatever details are added or deleted from the initial CRA, it will always leave the idea untouched that the appropriately programmed computer really is a mind.

Thus, instead of reflecting the logic contradictions of strong AI, Searle’s thought experiment circles around its own analogies and their explanatory status.

There is, as the 17th century example has shown, only one possible outcome of deconstructive machine-related thought experiments. One thought experiment will be followed by a next, more detailed one, which is followed by another one, and so on and so forth, ad infinitum. [9] And this is indeed pretty much what happened in the few decades that followed Searle’s essay.

Outlook

For future research in the field of Cognitive Science, the previous analysis of the CRA seems to suggest the following meta-guidelines.

1. Any theory of human or machine intelligence should, as a minimum requirement be self-aware of the thought experiments introduced and their role as constructive heuristic aids. Their role in theory making and their reference to our implicit set of “background knowledge” and 'pre-established' analogies should be stated as explicitly as possible. Special care should be taken in

The deconstructive part in the CRA

Back to the Chinese Room argument. Searle seems to have tried to introduce a machine-based thought experiment for the purpose of deconstructing a theory. And that is, as the examples from the 17th century have demonstrated, an undertaking bound to fail.

pointing out the scope and overall significance of common sense notions within the overall argument.

2. With thought experiments that feature machine settings, it should be stated explicitly that the outcome must always be constructive, and therefore will have limited explanatory use.

3. Deconstructive machine-related thought experiments are from an epistemological point of view, circular and irrelevant to theory. They are of little use in terms of guiding future research.

REFERENCES