Khashayar Etemadi
April, 2021

Meaning of understanding in cognitive science: The Turing test and the Chinese Room Argument

---

Turing Test [1] is a thought experiment in defense of the claim that “machines can think”. Its setup is as follows. There is a human in room A, and a machine in room B. There is also an interrogator in room C. The interrogator does not know if the human is in room A or B. So, she asks several questions from each room and gets written answers to decide in which room the machine sits and in which one the human sits. The machine tries to produce human-like answers in order to mislead the interrogator. The question is: “is it possible to have a machine that answers questions so much like humans that the interrogator cannot reliably determine whether the human is at room A or B?”

Alan Turing argues that if the answer to the question above is “yes”, it means machines can think. This argument is based on the behaviorist view about mind, which claims “there is no knowable difference between two states of mind (beliefs, desires, etc.) unless there is a demonstrable difference in the behavior associated with each state”. This entails that if a machine behaves so similar to humans that now one can demonstrate the difference between them, there is no knowable difference between the machine’s state of mind and a human’s state of mind. If the behaviorist view is true, Turing’s argument is sound as well, because if we can have a machine that makes it impossible for the interrogator to find the difference between it and a human being, this machine has the same knowable state of mind as humans, in other words, “it can think”.

At first glance, the behaviorist view and thus Turing’s argument seems to be sound. However, it has been subject to many criticisms. The Chinese Room Argument is one of these criticisms [3].

Similar to Turing’s argument, the Chinese Room Argument starts with a thought experiment as follows. There is an american person in a room who does not know anything about the Chinese language. This person is given a set of comprehensive instructions. The instruction is a long list that maps all possible questions in Chinese to their answers in English. The questions are written in Chinese and the answers are written in English. A person outside the room who does not know the person inside the room, asks Chinense written questions from the person inside. The person inside the room looks at the instructions, finds the question and copies the correct answer to it and gives the answer back to the person outside the room. The result is that even though the person inside the room does not understand Chinese, the person outside the room cannot distinguish between his/her answers and a native Chinese person’s answers.

The Chinese Room Argument argues that the Chinese Room experiment shows even if there is a machine that passes the Turing test, it does not mean that machine can

understand the questions just as the wo/man in the Chinese Room does not understand Chinese questions.

I personally agree with the Chinese Room Argument. It is a sharp and sound argument against Turing's argument.

Conclusion:

The Turing's and the Chinese Room arguments in favor of and against the claim that "machines can understand", respectively. They are related to artificial intelligence's understanding ability, not scientific understanding that we discussed in our lecture session. With regard to these two arguments, I vote for the Chinese Room Argument: passing the Turing test does not entail ability to understand.

References:

[1] https://en.wikipedia.org/wiki/Turing_test
[2] https://plato.stanford.edu/entries/behaviorism/
[3] https://plato.stanford.edu/entries/chinese-room/