Searle against AI

Jeff Speaks
PHIL 30304

November 29, 2018

1 Weak vs. strong AI

Searle begins by distinguishing weak vs. strong artificial intelligence (AI).

The claim that weak AI is possible is the claim that machines could simulate the behavior of beings like us which have mental states.

One test for whether a machine exhibits weak AI is what Alan Turing called 'the imitation game,' which has come to be known as the 'Turing test.'

The claim that strong AI is possible is the claim that machines could have the kinds of mental states we have.

We can further distinguish between two versions of strong AI: the claim that machines can think, and the claim that machines can enjoy conscious states.

It is sometimes claimed that passing the Turing test is also sufficient for strong AI. Counterexamples: Blockhead, zombies.

Searle's aim is to argue against the possibility of strong AI, focusing on the first version of that claim.

2 The Chinese room

Searle argues against the possibility of strong AI via the following thought experiment:

> 'Imagine a native English speaker who knows no Chinese locked in a room full of boxes of Chinese symbols (a data base) together with a book of instructions for manipulating the symbols (the program). Imagine that people outside the room send in other Chinese symbols which, unknown to the person in the room, are questions in Chinese (the input). And imagine that by following the instructions in the program the man in the room is able to pass out Chinese symbols which are correct answers to the questions (the output). The program enables the person in the room to pass the Turing Test for understanding Chinese but he does not understand a word of Chinese.
>
> ... The point of the argument is this: if the man in the room does not understand Chinese on the basis of implementing the appropriate program for understanding Chinese then neither does any other digital computer solely on that basis because no computer, qua computer, has anything the man does not have.'

Here is one way to think about the argument:

1. The person in the room does not understand Chinese.
2. If the person in the room does not understand Chinese, then no computer can understand Chinese.
3. If no computer can understand Chinese, then no computer can understand any language.
4. If no computer can understand any language, then no computer can think.

C. No computer can think.

3 The system reply

One central response to Searle's argument denies (2). On one way of pushing this objection, the man in the room does not understand Chinese, but the entire system — including the instructions, etc. — does. The man is after all just part of the relevant machine (he plays something like the role of a CPU.)

Searle's main reply is to imagine a case in which the man memorizes all of the instruction books. Then the man is the machine — but still, Searle thinks, would not understand Chinese.