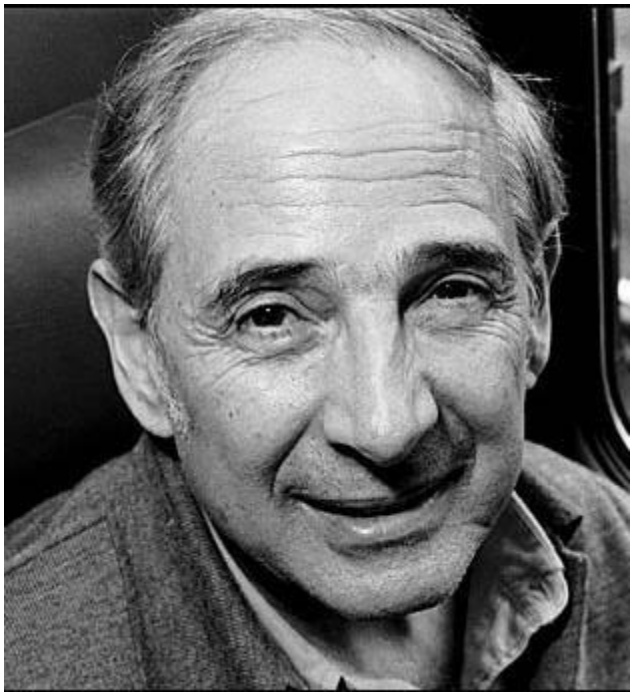

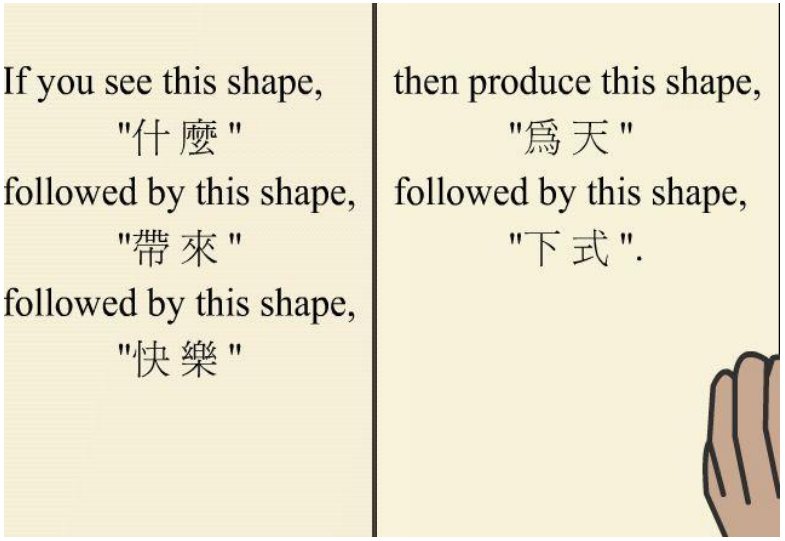

Searle's Chinese Room

Do chatbots *understand* what they're saying?

Functionalism and AI

- AI (Artificial intelligence) tries to design computer programs that will perform mental tasks of some kind.

- The whole idea of AI assumes functionalism. Functionalism says that the mind is software, not hardware.

- (In his “Chinese Room” paper, Searle attacks functionalism.)

Weak AI:

The programmed computer *simulates* the mind. It is a research tool, for testing psychological explanations. It can perform tasks that require thought in humans.

Strong AI:

The programmed computer *is* a mind. The computer really *understands* and is *conscious*.

"Mr. Lemoine caused a stir last month when he told *The Washington Post* that he believed Google's LaMDA, was sentient — unleashing fears that A.I. was moving closer to a dystopian sci-fi film and a raucous debate over whether a computer program can really have a soul." (*NYT*, July 2022)

- No one questions the possibility of weak AI, but strong AI is controversial.

- Searle focuses on ‘intentionality’ in this paper.
  - But *conscious* intentionality is probably an even bigger problem for computer programs.
- Intentionality = meaning, significance, or “aboutness”, understanding.
  - Thus ‘intentional’ mental states *represent* external states of affairs.

Intentionality

- Sentences and images can have intentionality, but only by association with a mental state.
- Mental states have *intrinsic*, or 'original' intentionality.

"the cat is next to a ball"

Do chatbots have intentionality?

- Searle doubts that computer *programs* can have intentionality in this sense.
- He discusses a chatbot program coded by Roger Shank, that can answer questions about restaurants.
  - (This is very different from today's LLMs like ChatGPT.)
- Does the chatbot *understand* what it's saying?
  - Does it have *original* intentionality?

How chatbots work

script

general information that the virtual person has (e.g. people won't usually eat a badly burned hamburger).

story

e.g. "the waiter brings a man a burned hamburger and he storms out without paying"

program

Some *very* complicated rules, based on dissecting the sentences, seeing formal (structural) relationships, reordering words, substituting terms, etc.

Program instructions are things like:

To any question of the form: **“Do you like [X]?”** reply:

“**Yes, I like [X]**” if X is a member of the set {chocolate, money, fast cars, philosophy, …}, but reply

“**No, I don’t like [X]**” if X is a member of {the smell of a wet dog, watching *Glee*, …}, and reply

“**I’m not sure, I don’t know what [X] is**” if X is on neither list.

questions

e.g. “Did the man eat the hamburger?”

responses

e.g. “No, he didn’t eat the hamburger”

- Suppose that the chatbot's answers to the questions are convincing, as good as those of a real English speaker. *Does the computer understand what it's saying?*

- No. Not a word of it, says Searle.

- Is Searle right about this?

The Chinese Room

- Searle's argument that chatbots lack original intentionality involves a thought experiment about the 'Chinese Room'.

- The basic idea is that a computer program implements a certain *function*, in the mathematical sense – it converts inputs into outputs.
- According to AI (and 'functionalism'), creating a mind is simply a matter of creating the right function.

Chinese Room Thought Experiment

- To show this, Searle imagines that *he himself* does the job of the computer, obeying the chatbot program's commands.

- Searle is in a room with a 'script', a 'story', some 'questions' and a 'program'.

- The trick is that the script, story, questions and answers are all in *Chinese*, a language that Searle doesn't speak at all. (The program is in English, so Searle understands that.)

If you see this shape,
"什 麼 "
followed by this shape,
"帶 來 "
followed by this shape,
"快 樂 "

then produce this shape,
"爲 天 "
followed by this shape,
"下 式 ".

- Suppose that the answers to the questions are convincing, as good as those of a real Chinese speaker. *Does Searle understand Chinese?*

- No. Not a word of it, says Searle. He has no idea what any of the questions, or his answers, mean. He is just cutting and pasting symbols, according to rules.

“1. As regards the first claim, it seems to me quite obvious in the example that I do not understand a word of the Chinese stories. I have inputs and outputs that are indistinguishable from those of the native Chinese speaker, and I can have any formal program you like, but I still understand nothing.” (p. 351)

- In other words, meaning does not arise from running a program, no matter how sophisticated. (Searle argues)

Clarifications

1. What *is* "understanding" anyway? More than a mere *function*, says Searle. When a thermostat turns the furnace off, does it think "it is too hot in here"?
   - No. (Certainly not *conscious* intentionality.)

2. Could a machine think? Yes, says Searle, *we are such machines*. But not in virtue of the *program*, he thinks. The hardware of the machine is relevant.

- “Yes, but could an artificial, a man-made machine, think?” Assuming it is possible to produce artificially a machine with a nervous system, neurons, with axons and dendrites, and all the rest of it, sufficiently like ours, again the answer to the question seems to be obviously, yes. If you can exactly duplicate the causes, you could duplicate the effects. And indeed it might be possible to produce consciousness, intentionality, and all the rest of it using some other sorts of chemical principles than those that human beings use. It is, as I said, an empirical question.”

  (p. 352)

“OK, but could a digital computer think?”

If by “digital computer” we mean anything at all that has a level of description where it can correctly be described as the instantiation of a computer program, then again the answer is, of “yes”, we are the instantiations of any number of computer programs, and we can think.”

(p. 353)

- “But could something think, understand, and so on solely by virtue of being a computer with the right sort of program? Could instantiating a program, the right program of course, by itself be a sufficient condition of understanding?” This I think is the right question to ask, though it is usually confused with one or more of the earlier questions, and the answer to it is no.” (p.353)

- **Functionalism**: If two systems are functionally equivalent (same outputs for the same inputs) then they're mentally equivalent (same consciousness, intentionality, etc.)

- *Searle opposes functionalism*, not materialism. A complex system of water pipes, etc. cannot be conscious, even if it executes the right program, Searle thinks.

What are the pipes thinking about?