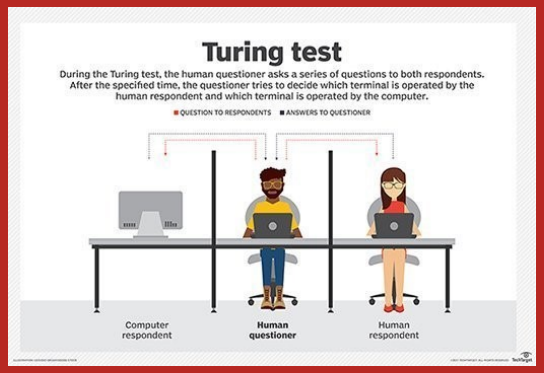
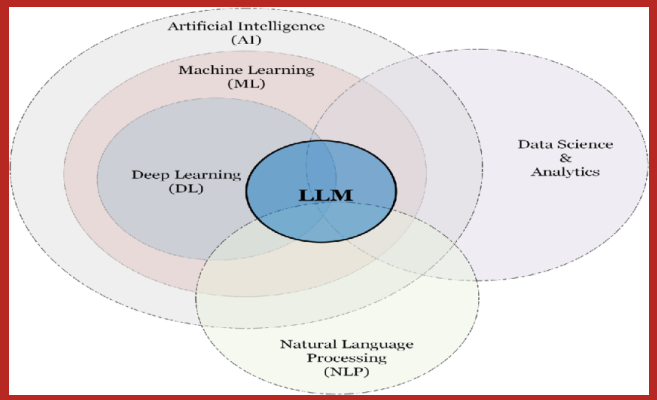
The Nature of intelligence and philosophy of mind in the age of Large Language Models



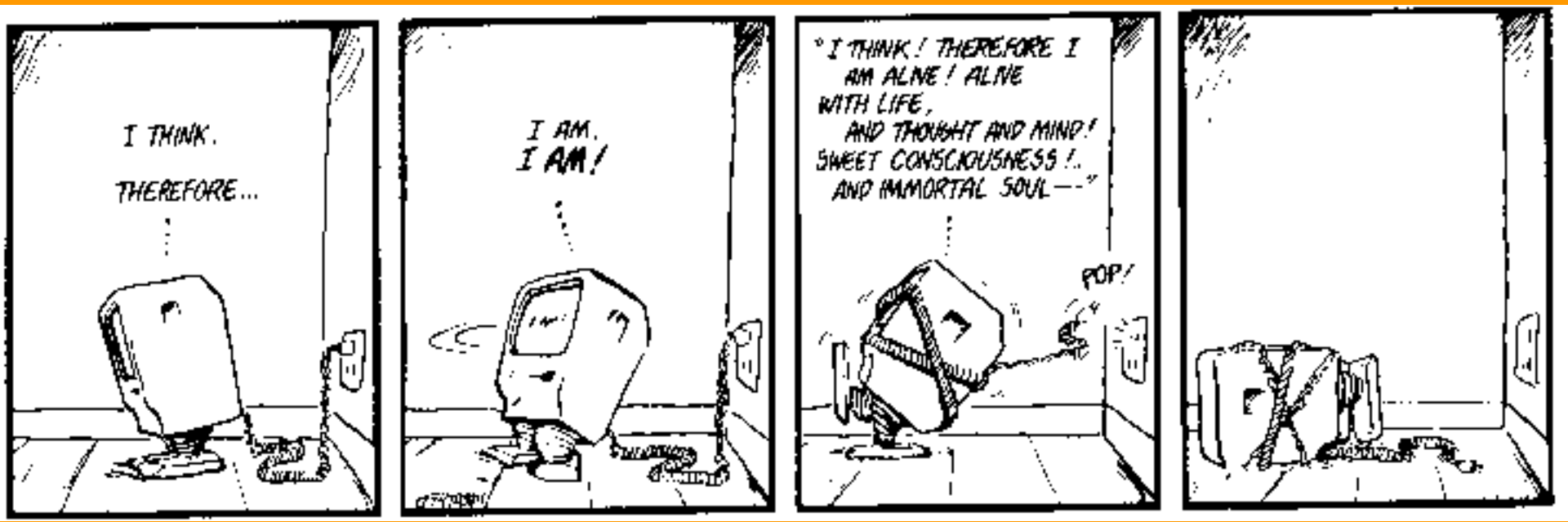


Luca Consoli and Michel Vitale, 16-12-2024

Outline

- Class exercise (Michel)
- The Chinese Room Argument in the time of GOFAI
- The Chinese Room Argument in the time of LLM’s

Bloom County on Strong AI



THE CHINESE ROOM

- **Searle's target: "Strong AI"**
  - **An appropriately programmed computer *is* a mind—capable of understanding and other propositional attitudes**

The Gedankenexperiment

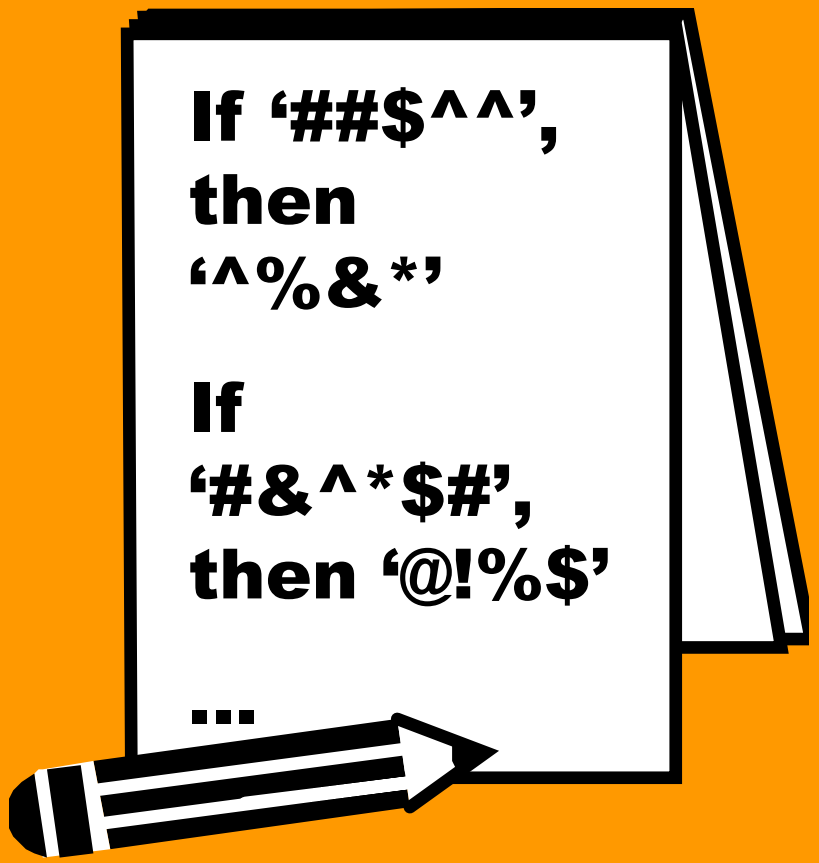
- **Searle, who knows no Chinese, is locked in a room with an enormous batch of Chinese script.**

The Gedankenexperiment

- **Slips of paper with still more Chinese script come through a slot in the wall.**

The Gedankenexperiment

- Searle has been given a set of rules in English for correlating the Chinese script coming through with the batches of script already in the room.



The Gedankenexperiment

- Searle is instructed to push back through the slot the Chinese script with which the scripts coming in through the slot are correlated according to the rules.

The Gedankenexperiment

- But Searle, remember, knows no Chinese; he identifies the scripts coming in and going out on the basis of their shapes alone. And following the rules requires only this ability.

The Gedankenexperiment

- Suppose that those outside the room call the scripts going in ‘the questions’, the scripts coming out ‘the answers’, and the rules that Searle follows ‘the program’.

The Gedankenexperiment

- Suppose also that the program writers get so good at writing the program and Searle gets so good at following it that Searle's answers are indistinguishable from those of a native Chinese speaker.

The result

- **It seems clear that Searle nevertheless does *not* understand the questions or the answers; he is as ignorant as ever of Chinese.**

"Neih hou ma?"

The result

- **But Searle is behaving just a computer does, “performing computational operations on formally specified elements”**

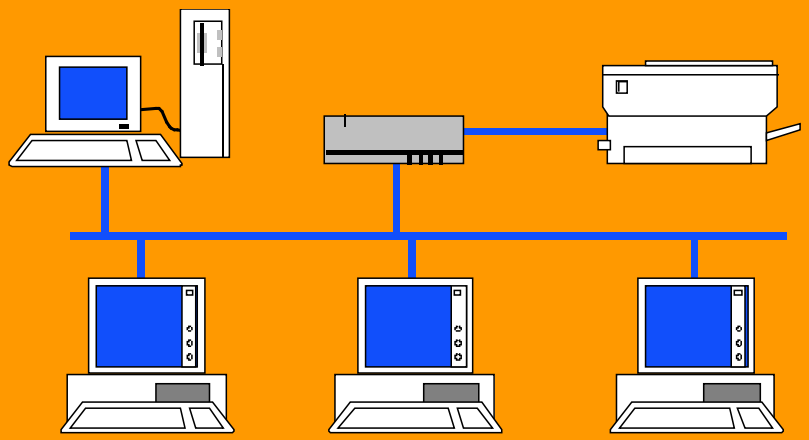
The result

- **Hence, manipulating formal symbols—which is just what a computer running a program does—is not sufficient for understanding or thinking.**

On the way to the replies

- **‘Understanding’—it comes in degrees, it’s vague…**
- **Searle: But I’m a *clear case* of understanding English sentences, and a *clear case* of failing to understand Chinese ones.**

The systems reply



- Of course the guy in the room doesn't understand. But the guy is just part of a larger system which does understand.

Systems reply: Searle's 1st rejoinder

- Let the guy (Searle) internalize all the elements of the system.
- But the guy still doesn't understand and neither does the system because there is nothing in the system that is not in him—the system is just a part of him.

Systems reply: Searle's 2nd rejoinder

- The idea is that, although the guy doesn't understand Chinese, somehow the *conjunction* of the guy plus the rulebook, the batches of script, the slot, and the room does understand. That's ridiculous.

Systems reply: Searle's 2nd rejoinder

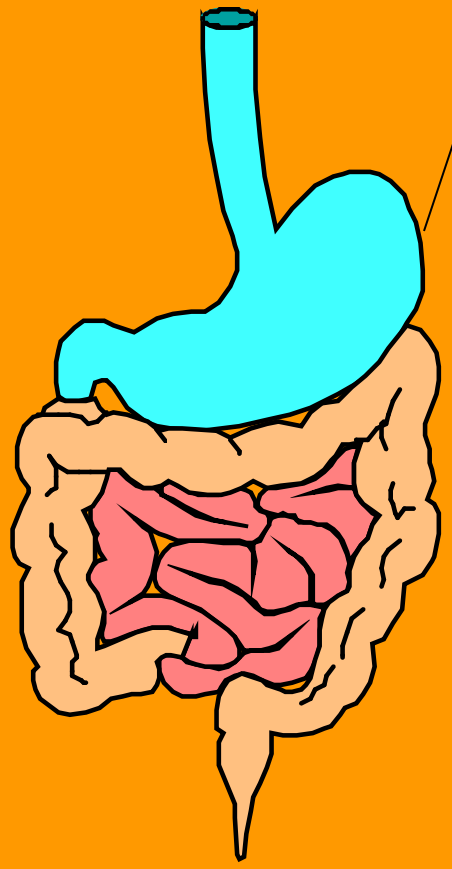
- Knowing what 'Do you want a hamburger?' means.
- One necessary condition: You have to know that 'hamburger' refers to hamburgers.
- How on earth does the Chinese Room system come to know facts like these?

Systems reply: Searle's 3rd rejoinder

- If the system counts as cognitive simply because it has certain inputs and outputs and a program in between, then all sorts of non-cognitive systems are going to count as cognitive.

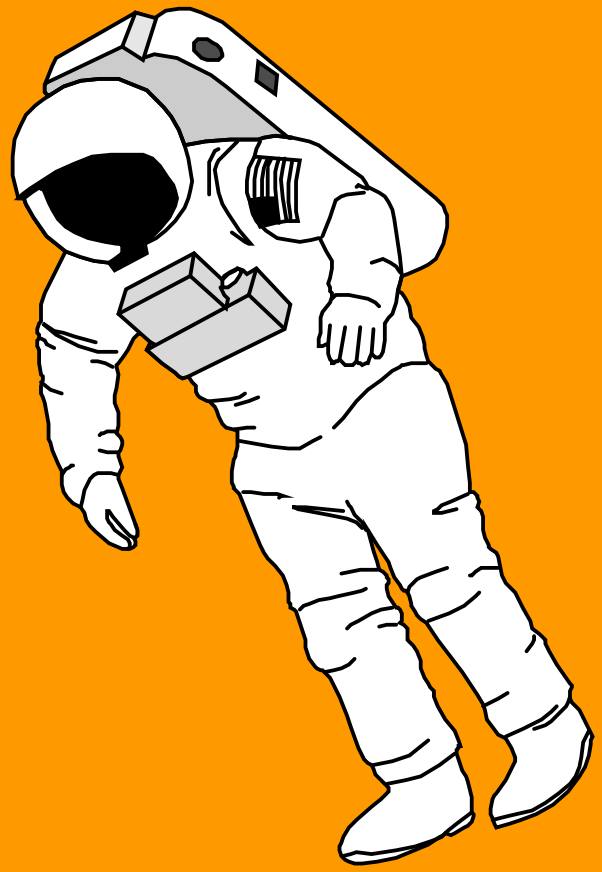
Systems reply: Searle's 3rd rejoinder

- Just an interesting result of the strong AI model?
- Can't then explain what *makes* the mental mental.



I understand too? Yippee!

The robot reply



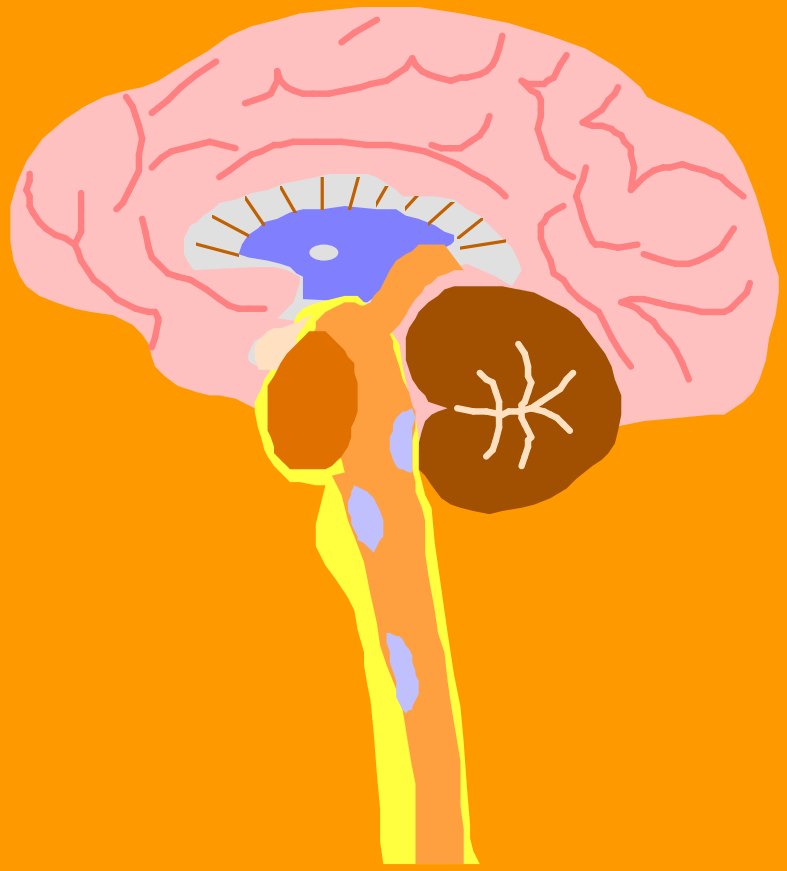
- **Put a computer in the head of a robot giving the computer "perceptual" and motor capacities—this will bestow understanding and intentionality.**

Robot reply: Searle's rejoinder

- Note that the reply concedes that manipulating formal symbols does not add up to understanding.
- Put the Chinese Room in the head of a robot. Still no understanding.

Brain simulator reply

- **Design a program that simulates the actual sequence of neuron firings that occur in a native Chinese speaker when he/she is understanding a story--wouldn't this do it?**

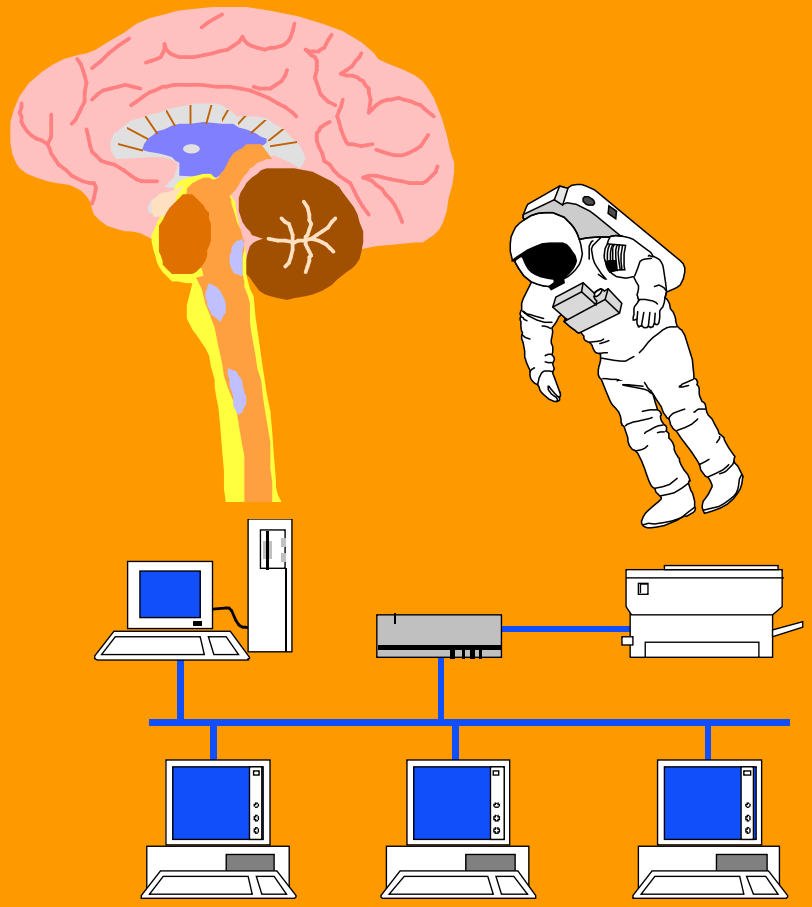


Brain simulator reply: Searle's rejoinders

- The reply concedes something important: that we might need to know how the brain works in order to know how the mind works. But this seems like an abandonment of Strong AI.
- Chinese plumbing in the Chinese room: Still no understanding.

Combination reply

- **Simulated brain in the head of a robot comprising a thinking system.**



Combination reply: Searle's rejoinder

- We would find it "irresistable" to attribute intentionality to such an entity. But that has entirely to do with its behavior. Once we learned how the entity functioned we'd withdraw our attributions.

Other minds reply

- If we refuse to attribute intentionality to the robot, won't we have to refuse to attribute intentionality to other people?
- Rejoinder: This confuses an epistemological problem with a metaphysical one.

Does Searle's argument hold with LLMs?

- **Pro: chinese Room argument still strong**

1. **Lack of true understanding**
   - LLMs simulate understanding without actual meaning (Searle)
   - Key features of human mental life could not be captured by formal rules for manipulating symbols (Dreyfus, who's critics of AI).
2. **Functionalism vs. understanding**
   - LLMs exhibit behavior, not true intentionality or consciousness (Metzinger)
3. **Symbol Grounding Problem**
   - LLMs lack sensory or intentional grounding of symbols (Andy Clark and his thoughts around microfunctionalist accounts of consciousness and mentality).

Does Searle's argument hold with LLMs?

- **<u>Against: Chinese Room Argument Weak</u>**

1. **Scale and Complexity**
    - LLMs generate complex, meaningful responses, suggesting understanding (Dennett thinks future AI could evolve in some way and get out of hands…)
2. **Emergent Properties**
    - LLMs display behaviors that resemble understanding, despite individual components lacking it (Chalmers).
3. **Turing Test and Pragmatism**
    - In a more "pragmatic" sense, if LLMs pass the Turing Test, they can be considered to "understand" (following Turing…)

The Octopus Test (Bender & Koller, 2020)

"Say that A and B, both fluent speakers of English, are independently stranded on two uninhabited islands. They soon discover that previous visitors to these islands have left behind telegraphs and that they can communicate with each other via an underwater cable. A and B start happily typing messages to each other.

Meanwhile, O, a hyperintelligent deep-sea octopus who is unable to visit or observe the two islands, discovers a way to tap into the underwater cable and listen in on A and B's conversations. O knows nothing about English initially but is very good at detecting statistical patterns. Over time, O learns to predict with great accuracy how B will respond to each of A's utterances.

The Octopus Test (Bender & Koller, 2020)

Soon, the octopus enters the conversation and starts impersonating B and replying to A. This ruse works for a while, and A believes that O communicates as both she and B do — with meaning and intent. Then one day A calls out: "I'm being attacked by an angry bear. Help me figure out how to defend myself. I've got some sticks." The octopus, impersonating B, fails to help. How could it succeed? The octopus has no referents, no idea what bears or sticks are. No way to give relevant instructions, like to go grab some coconuts and rope and build a catapult. A is in trouble and feels duped. The octopus is exposed as a fraud." (https://kottke.org/23/03/the-octopus-test-for-large-language-model-ais)