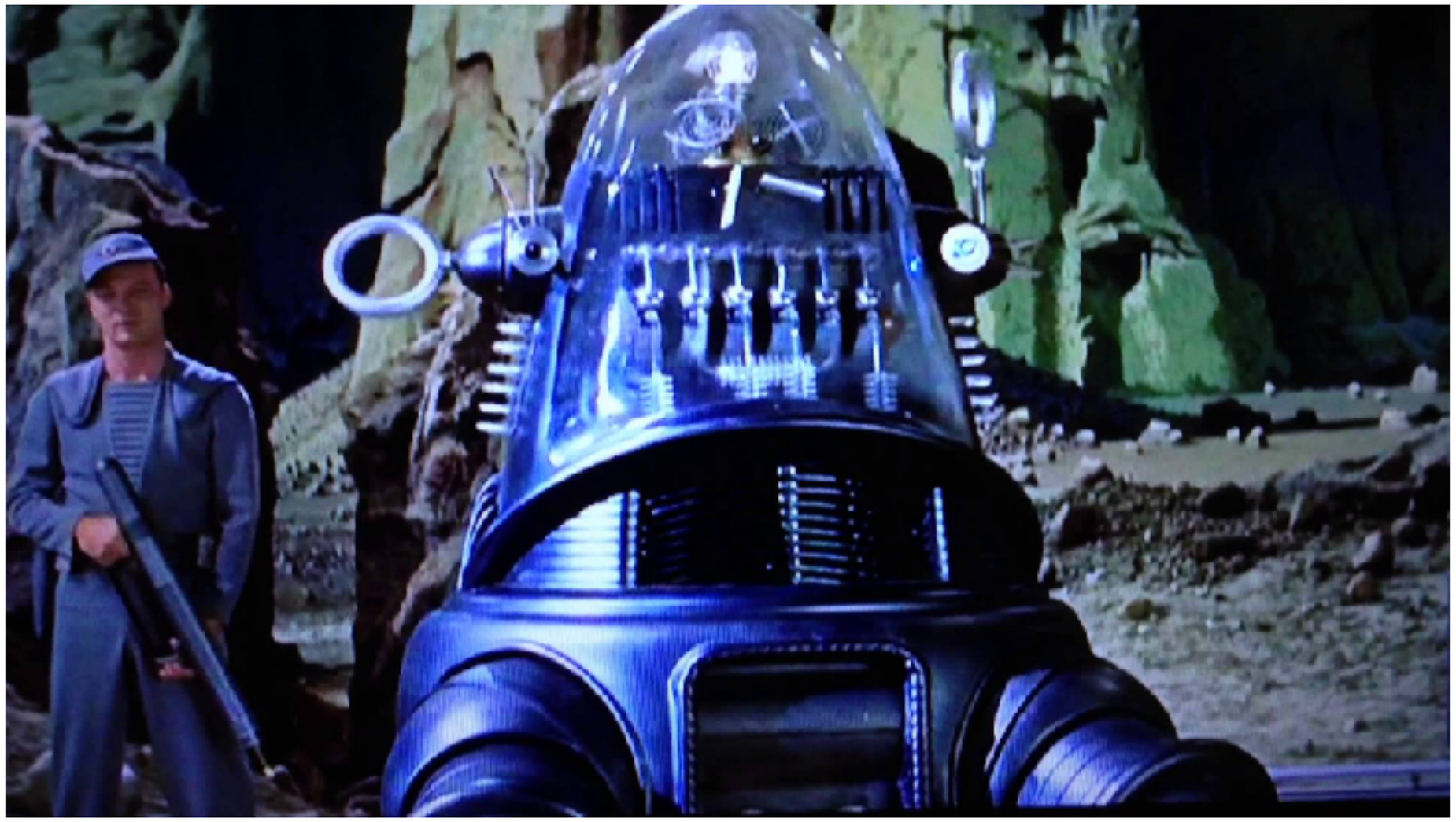


Robby the robot in Forbidden Planet, 1956

Mental representations and symbol systems

Functionalism II

References

- Alan Newell and Herbert Simon, « Computer science as empirical enquiry », 1975.
- Jerry Fodor, The language of thought, 1975.
- Jerry Fodor, « Why there has still to be a language of thought », 1987
- Jerry Fodor, LOT 2, 2008.
- Alan Turing, « Computing machinery and intelligence », 1950.
- John Searle, « Minds, brains and programs » (with peers commentaries), 1980.
- For a synthetic presentation: José Bermudez, *Cognitive Science: An introduction to the science of the mind*, Cambridge, 2010.

Plan

- Fodor and the langage of thought;
- Searle’s Chinese room argument;
- Discussion of Searle’s argument.

I From AI to the language of thought

The physical symbol system hypothesis

- In 1975 the *Association of Computing Machinery* gave its Turing award to Newell and Simon, two forefathers of classical AI.

- In their lecture, they put forward the following thesis:

    - **A physical symbol system has the necessary and sufficient means for general intelligent action.**

- Necessary condition ==> humans beings are « physical symbol systems » since they are intelligent agents;

- Conversely: there is no obstacle in principle to constructing an artificial mind, provided that one solves the problem of constructing **a physical symbol system**.

« A physical symbol system consists of **a set of entities, called symbols, which are physical patterns that can occur as components of another type of entity called an expression (or symbol structure)**. Thus a symbol structure is composed of a number of instances (or tokens) of symbols related in some physical way (such as one token being next to another). At any instant of time the system will contain a collection of these symbol structures. **Besides these structures, the system also contains a collection of processes that operate on expressions to produce other expressions**: processes of creation, modification, reproduction, and destruction. A physical symbol system is a machine that produces through time an evolving collection of symbol structures. »

The classical AI conception of the mind

- Token symbols are physical patterns ==> human thoughts can be interpreted as token symboles implemented in brain states.

- Atomic symbol types can be combined to give more complex symbol structures, just like in a formal language, atomic symbols can be combined by syntactic rules.

- A physical symbol system contains rules for manipulating these complex symbol structures.

- These rules can themselves be represented by symbol structures: cf. Turing Machines, or actually any kind of programming language.

The LOT hypothesis

- « When I wrote LOT 1 I remember thinking (and not without a certain satisfaction) that it didn't contain a single new idea. I believed (quite wrongly, it seems to me in retrospect) that **my enterprise was merely journalistic**: I was writing an exposition of what I took to be an emerging, interdisciplinary consensus about how the mind works; the theory that was then just beginning to be called 'cognitive science' (…) a cognitive science that was mentalistic rather than behavioristic. Fodor, *LOT 2:* (p. 3-4).

- Core assumption:

  - The basic symbol structures in the mind that carry information are **token sentences in an internal language of thought**.

  - Information processing works by **transforming those sentences into other sentences: information processing = symbol manipulation.**

The arguments from intentional (cartesian) realism

- Some mental states can be described as relations to propositions.

    - For instance: Believing that Camberra is the capital of Australia = a Belief-relation with the proposition <Camberra is the capital of Australia>.

- These states are real, and have real causal powers.

    - Suppose that I want to visit the Capital of Australia, and that I know that Camberra is the Capital of Australia. These propositional attitudes can casually explain my buying a flight ticket to Camberra.

- NB: Intentional realism is opposed both by eliminativists and by instrumentalists, who consider that propositional attitudes are fictional entities.

- One of Fodor's starting point is the recognition that belief/desire folk psychology is very successful;

- On his view, the best explanation of its success is its truth;

- Which means that Fodor is an intentional realist: he thinks that some of our mental states have intentional content, and that these states can cause behaviors.

- Problem: how can intentional states have causal powers?

- The LOT hypothesis can be seen as an answer to this question:

    - Symbols have intentional properties;

        - For instance, « Dogs » denote dogs;

    - Physical systems can manipulate symbols, therefore symbols can cause behavior (via syntactic manipulations);

    - Hence, by inference to the best explanation, thoughts are symbol structures!

RTM: the representational theory of mind

- « Tokens of cognitive mental states are tokens of relations between creatures and their mental representations. »

- « Tokens of mental processes are ‘computations’; that is, causal chains of (typically inferential) operations on mental representations. »

    - To this extent, classical AI is right to see mental states tokens as formulas in a language-like medium!

- What classical AI missed, according to Fodor: « truism: that the mind’s main concern is not acting but thinking, and that paradigmatic thinking is directed to ascertaining truths. What minds do is think **about** things. »

- So symbols are essential to thought, yes …

- … but symbols' relations to the objects in the world that they represent are essential, because without those relations symbols wouldn't have any content …

- … and content is essential too: that's because they have content that thoughts cause behavior. Consider this practical syllogism from LOT2 (p. 15):

1. Thinks: I want P

2. Thinks: P only if Q

Conclusion : Acts to bring it about that Q

- What connects the two premises in this reasoning is their overlapping contents: that's because P is both the intentional object of a desire and of a belief that the action is caused.

A semantic engine driven by syntax

- AI teaches us that computational rules manipulate symbols according to their **syntactic** properties;

- On the other hand, folk psychology teaches us that rational transitions between thoughts is determined by their **semantics**;

- If follows that causal transitions between sentences in the language of thought have to respect the rational relations between the contents of those sentences in the language of thought.

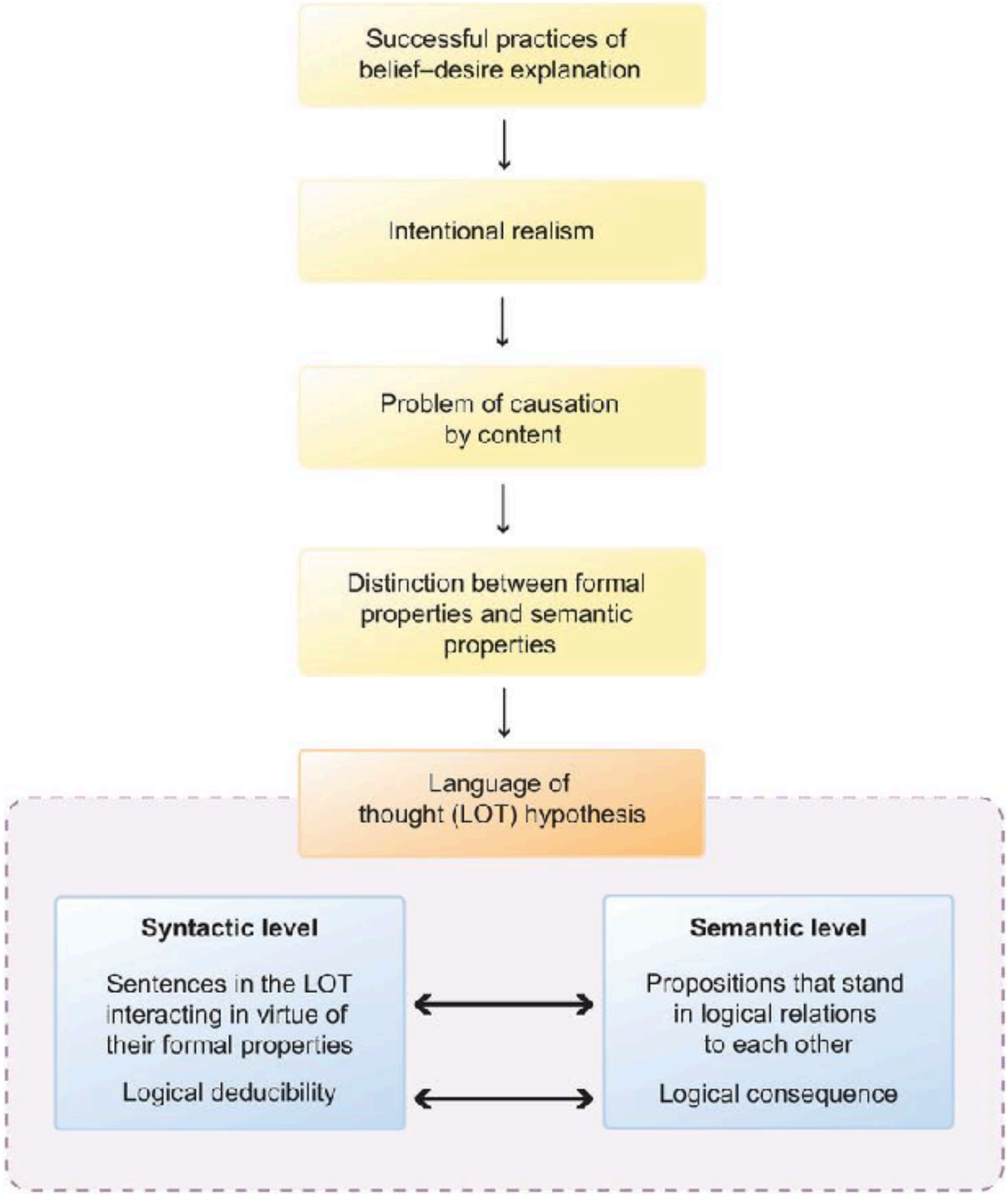
    - Cf the completeness theorem for first-order logic.

    - To each valid argument corresponds a proof, which is a syntactic object ==> the semantic consequences of axioms in FOL can always be computed by a machine.

- José Bermudez: « Fodor's basic proposal, then, is that we understand the relation between sentences in the language of thought and their content (or meaning) on the model of the relation between syntax and semantics in a formal system. » Cognitive Science: An Introduction to the Science of the Mind (p. 157).

Symbols	Transformation rule	Meaning
1. G*a*		1. Georgina is tall
2. F*a*		2. Georgina has red hair
3. (F*a* & G*a*)	If complex symbols “S” and “T” appear on earlier lines, then write “(S & T)”	3. Georgina is tall and has red hair
4. ∃*x* (F*x* & G*x*)	If on an earlier line there is a complex symbol containing a name symbol, then replace the name symbol by “*x*” and write“∃*x* —” in front of the complex symbol	4. At least one person is tall and has red hair

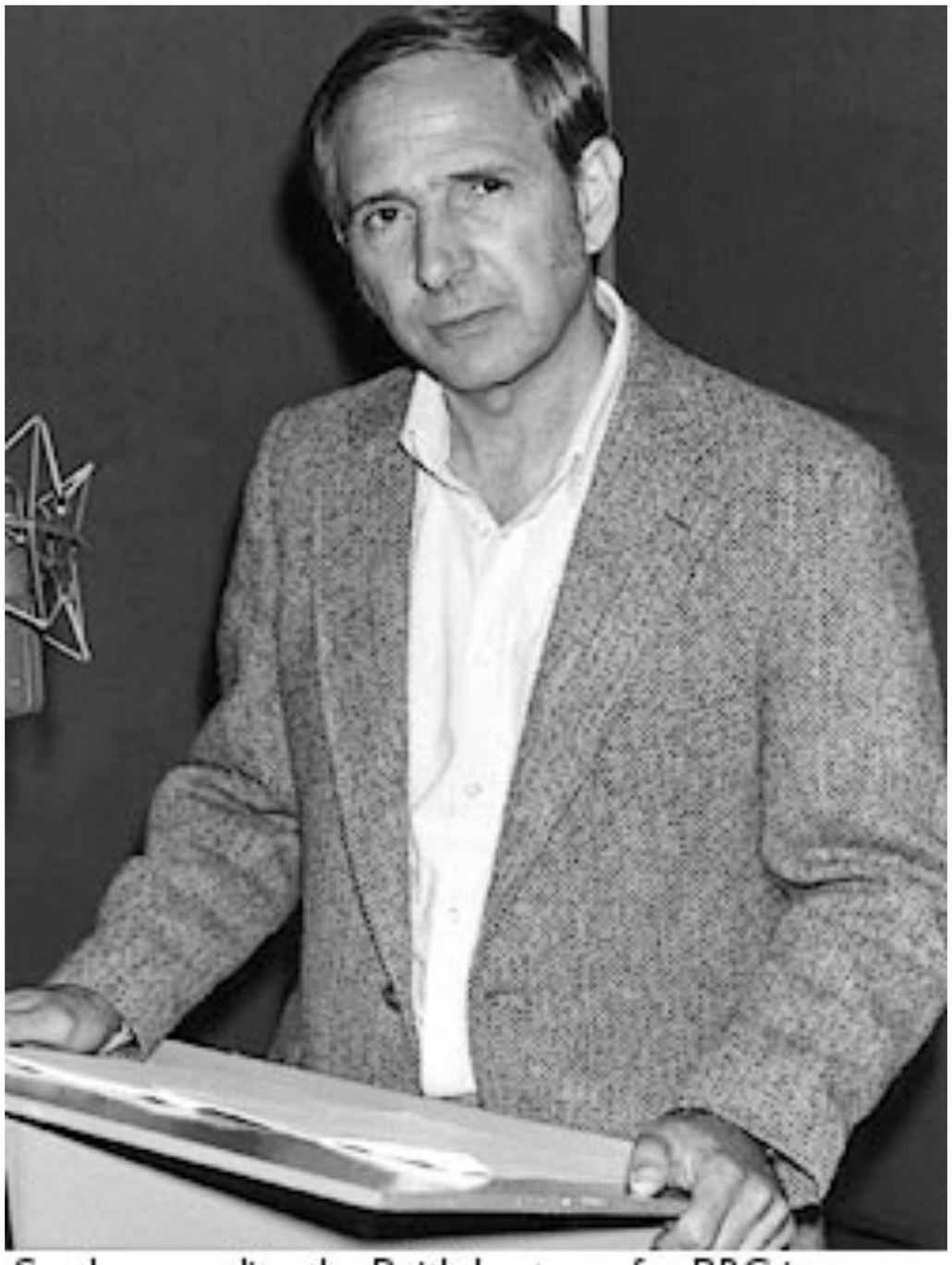
From : Bermudez, An introduction to the science of the mind, p. 158



Source: Bermudez

II The Chinese room argument

Searle vs the robots



Searle recording the Reith Lectures for BBC in London, summer of 1984, where he first presented the "Chinese Room" argument.

The Turing test

- Turing 1950 (Mind);

- The machine passes the test iff the evaluator cannot reliably tell the machine from a human;

- Replace the question « can machine think » by a simpler one: « Are there imaginable digital computers which would do well in the *imitation game*? »

- Eliza: 1966.

- Cf. Loebner's prize winner Mitsuku, or other chatbots, for more recent illustrations.

Joseph Weizenbaum, ELIZA's creator, on the Turing test:

« [In] artificial intelligence ... machines are made to behave in wondrous ways, often sufficient to dazzle even the most experienced observer. But once a particular program is unmasked, once its inner workings are explained ... its magic crumbles away; it stands revealed as a mere collection of procedures ... The observer says to himself "I could have written that". With that thought he moves the program in question from the shelf marked "intelligent", to that reserved for curios »

Searle’s thought experiment

- A person in a room receives pieces of paper through one window, with queries in Chinese written on them, and passes out written answers through another window.

- He doesn’t understand Chinese at all: the answers are determined by syntactic rules written in a huge instruction manual.

- The person in the room is the analogue of a Turing machine’s head: he applies syntactic rules to symbols in order to manipulate them, and the rules only depend on the symbols’ syntax, not on their semantic.


I'm just manipulating squiggles and squoggles to produce Chinese language behavior. But I don't understand Chinese. This rule book is in English.
在這屋裡的任何人或物,一定懂中文。
[Whoever or whatever is in that room is an intelligent Chinese speaker!]
Take a squiggle-squiggle sign from basket number 1 and put it next to a squoggle-squoggle sign from basket number 2.
1
2
3

- From a behavioral point of view, the Chinese room is behaving exactly like a Chinese speaker: it would pass Turing's test.

- Nobody understands Chinese in this situation however: neither the person in the room, nor the room itself.

- So syntactic formal manipulation cannot constitute understanding.

Comparison with the Turing test

- The Turing test is weaker than the symbol system hypothesis, because it is purely behavioral:
    - In order to pass the test, a machine just has to associate the right outputs to certain given inputs;
        - This does not even imply that it does it by manipulating symbols, nor that it is able to draw inferences;
            - Many chatbots thus give an illusion of understanding by using associations, pattern matching techniques without any reasoning abilities.
        - Searle’s argument, on the other hand, is tailored to refute the symbol system hypothesis, with the following argument:

Searle’s argument

- « Understanding a language, or, indeed, having mental states at all, involves more than having just formal symbols.

- It involves having an **interpretation**, or a meaning attached to those symbols.

- And a digital computer, as defined, cannot have more than just formal symbols because the operation of the computer, as I said earlier, can only be defined in terms of its ability to implement programs.

- And these programs are purely formally specifiable—that is, they have no semantic content. » (BBC talk)

- Understanding is a necessary condition for intelligence; computers cannot understand anything; conclusion: an artificial intelligence is impossible, unless it is implemented in a brain.

III Discussion

The system reply

- There is an equivocation in the argument:

- On the physical symbol system hypothesis, any sufficiently complex physical symbol system can display intelligent behavior;

- Searle in his argument does not deny that the system as a whole is intelligent: he denies that the person in the Chinese room—that is, a part of the system—, is intelligent.

- This is problematic because most computational functionalist would agree that many parts of complex symbol systems are not intelligent, even though intelligence emerges from the causal interactions of those parts.

Searle’s answer

- Imagine an internalized Chinese room; for instance, imagine that you have learned all the rules in the instruction book by heart. Do you think that you would understand Chinese?

- According to Searle you wouldn’t because you wouldn’t understand a world of Chinese.

- This is debatable: it is almost impossible to imagine someone having internalized a Chinese room, so our intuitions are very indeterminate.

The robot objection

- An advocate of this objection would agree with Searle that the Chinese room doesn’t understand Chinese, but she would not consider that this has anything to do with the relationships between syntax and semantic.

- She would insist that an intelligent robot has to be **embodied**: understanding is a very complex ability, which manifests itself in the way an agent linguistically interact with other agents.

- For being able to carry out such interactions, an agent has to be able both to perceive the world and to act in the world.

- For that, she must have sense organs and an action systems (limbs, and the capacity to move those limbs at will).

- Only a being-in-the-world can be an intelligent agent ==> on the whole a very Heideggerian objection!

Searle’s reply

- If transduction (conversion from the energy in the sensory stimulus to information) is understood as symbol-construction and symbol-manipulation, then the Chinese room objection in unscathed.

- Let us consider a robot able to perceive a stop traffic sign and to stop as a result. It would construct an inner symbol from the light energy, apply a rule to this symbol by formally manipulating it in its central system; this would lead up to sending a command to the action system, and the robot would stop.

- At no step in this symbolic manipulation process is there any instance of understanding, according to Searle.

The symbol grounding problem

- Upshot of the discussion: Searle's merit consists in having highlighted a very deep philosophical problem, **the symbol grounding problem**.

    - How do our thoughts get meaning or content? If we are physical symbol systems, how come the symbols instantiated in our brains are meaningful?

    - Searle's own solution to this problem is **biological**: only brain states can instantiate phenomenal properties or understanding; it follows that an artificial intelligence is metaphysically impossible

- This solution, however, may sound deeply dissatisfactory: why would brain states, and only brain states, instantiate intentionality? Why, and how, do consciousness and intentionality emerge from the brain?

Intentionality and the grounding problem

- Intentionality problem: how can we explain the fact that our mental representations **refer** to objects in the world or to aspects of situations?

- Grounding problem: how can we explain the fact that our representations (our words, our conscious experiences, our thoughts) are **meaningful to us, as subjects**?

- A lot of energy has been devoted to the solution of the intentionality problem by philosophers of mind.

    - For instance, according to Fodor, our mental representations refer because there are causal laws connecting the occurrences of mental symbols to their worldly referents.

- It is unclear, however, that a solution of the intentionality problem is enough to solving the grounding problem.

- Consider natural language:

    - In a sense, we know why words have the referents and the meanings they have: that's because we use them in certain ways (cf Wittgenstein);

    - This does not explain, however, why they are **meaningful to us** — a zombie having no access whatsoever to meanings could use words exactly like us.

- In the case of words, we could surmise that they are meaningful to us because our thoughts are meaningful, and because we can associate these meaningful thoughts to words uses.

- But then, why are thoughts meaningful in the first place? And what does it mean exactly that they are meaningful? Those are very deep open problems raised by Searle's thought experiment.