# Chinese room experiment

The Chinese Room Experiment is a philosophical experiment that was conducted by Searle in 1980 in order to understand whether the computer science advances of those years could be instances of the Strong AI hypothesis.

The Chinese experiment does not claim anything about the Weak AI assumption (it seems that Searle believes that it is possible to simulate intelligence) but from the experiment Searle concludes that it is not possible to show that a machine has a consciousness, understanding or a mind given that merely follows a set of rules written in a symbolic way. Remember that in those ages, the symbolic system hypothesis (GOFAI) was the mainstream assumption about intelligence. However, as we are going to see Searle shows that merely follow a set of symbols whose syntax is understood but not the semantics does not imply understanding.

In order to carry out the experiment, Searle exploits the program devised by Shank:

The aim of the program is to simulate the human ability to understand stories. It is characteristic of human beings’ story-understanding capacity that they can answer questions about the story even though the information that they give was never explicitly stated in the story. Thus, for example, suppose you are given the following story: "A man went into a restaurant and ordered a hamburger. When the hamburger arrived it was burned to a crisp, and the man stormed out of the restaurant angrily, without paying for the hamburger or leaving a tip." Now, if you are asked "Did the man eat the hamburger?" you will presumably answer, "No, he did not." Similarly, if you are given the following story: "A man went into a restaurant and ordered a hamburger; when the hamburger came he was very pleased with it; and as he left the restaurant he gave the waitress a large tip before paying his bill," and you are asked the question, "Did the man eat the hamburger?" you will presumably answer, "Yes, he ate the hamburger." Now Schank’s machines call similarly answer, questions about restaurants in this fashion. To do this, they have a "representation" of the sort of information that human beings have about restaurants, which enables them to answer such questions as those above, given these sorts of stories. When the machine is given the story and then asked the question, the machine will print out answers of the sort that we would expect human beings to give if told similar stories. Partisans of strong AI claim that in this question and answer sequence the machine is not only simulating a human ability but also (1) that the machine can literally be said to understand the story and provide the answers to questions, and (2) that what the machine and its program do explains the human ability to understand the story and answer questions about it.

Thus, strong AI advocates believe that Shank’s program is an instance of the Strong AI hypothesis and as such the program literally understands the stories.

The Chinese room experiment description is the following:

* Let’s suppose a person is in a room.
* Let’s suppose this person does not know Chinese at all but knows English.
* Let’s suppose that the Shank’s program can answer questions in Chinese to stories written in Chinese as well.

The person is given a story in Chinese and a set of questions in Chinese as well. In addition, he/she is given a set of instructions in English which resembles the Shank’s program code.

Thus, assuming that the Shank’s program works correctly, the person that does not know anything about Chinese correctly answers the questions following the instructions provided.

Let’s suppose that the same person conducts the same experiments but in English.

What is the difference between the Chinese experiment and the English experiment?

The difference is that while in the latter the Person really gain an understanding of the story and consequently can answer the questions, in the former the person is actually merely following a set of instructions as the computer does gaining no understanding of the story or the answer being provided.

In other words, the person is behaving as an instance of the Shank’s program.

More generally, instead of specifically talking about the Shank’s problem we can state the following:

Any person can follow the rules of a program code without understanding. Therefore, the merely following the rules of a computer program is not sufficient to state that an AI Agent is an instance of the Strong AI hypothesis as the agent could follow the rules in a similar way as a person and gaining no real understanding.

<http://web.archive.org/web/20071210043312/http://members.aol.com/NeoNoetics/MindsBrainsPrograms.html>

<https://plato.stanford.edu/entries/chinese-room/#ChinRoomArgu>

We can summarise the Seale’s argument as follows:

Let L be a natural language, and let us say that a “program for L” is a program for conversing fluently in L. A computing system is any system, human or otherwise, that can run a program.

1. If Strong AI is true, then there is a program for Chinese such that if any computing system runs that program, that system thereby comes to understand Chinese.
2. I could run a program for Chinese without thereby coming to understand Chinese.
3. Therefore Strong AI is false.

The first premise is based on the Strong AI assumption. Indeed, if a program that converses in Chinese exists and it is an instance of the Strong AI hypothesis then the actual program is a mind and not a simulation of the mind. In other words, the program reasons as a person would do. Thus, any person that reads the program can come to understand Chinese as it follows rules that are the same ones that his mind would follow. Therefore, the person can understand Chinese and learn it through the program.

The second premise tells that it is possible to follow the instructions of such a program but still not gaining any understanding of Chinese. This is supported by the Chinese experiment.

The conclusion is that the Strong AI assumption must be false as it is possible to follow a program that is an instance of the Strong AI assumption but still not gaining any understanding. Therefore, the program cannot be an instance of the AI assumption.

The argument above is an argument by contradiction.

## The larger philosophical issues as stake

Searle believes the Chinese Room argument supports a larger point, which explains the failure of the Chinese Room to produce understanding. Searle argued that programs implemented by computers are just syntactical. Computer operations are “formal” in that they respond only to the physical form of the strings of symbols, not to the meaning of the symbols. Minds on the other hand have states with meaning, mental contents. We associate meanings with the words or signs in language. We respond to signs because of their meaning, not just their physical appearance. In short, we understand. But, and according to Searle this is the key point, “Syntax is not by itself sufficient for, nor constitutive of, semantics.” So although computers may be able to manipulate syntax to produce appropriate responses to natural language input, they do not understand the sentences they receive or output, for they cannot associate meanings with the words.

Searle (1984) presents a three premise argument that because syntax is not sufficient for semantics, programs cannot produce minds.

1. Programs are purely formal (syntactic).
2. Human minds have mental contents (semantics).
3. Syntax by itself is neither constitutive of, nor sufficient for, semantic content.
4. Therefore, programs by themselves are not constitutive of nor sufficient for minds.

The third premise is supported by the Chinese room experiment. The conclusion is also the following:

* Programs are not minds

A consequence of the argument above is called Biological Naturalism.

## Biological Naturalism (Searle)

According to which mental states are high-level emergent features that are caused by low-level physical processes in the neurons, and it is the (unspecified) properties of the neurons that matter. Thus, mental states cannot be duplicated just on the basis of some program having the same functional structure with the same input–output behaviour; we would require that the program be running on an architecture with the same causal power as neurons.

In other words, according to biological naturalism it is not possible for a machine to have a real mind even though it has the same input-output behaviour (given a certain input outputs something like a real mind would do) because mental states rely on some properties of the neurons. That is, according to Biological Naturalism, the mind is strictly related to the brain and so it is not possible to recreate a mind without a brain.