John Searle's ‘Chinese Room’

This is an ideal experiment proposed by the philosopher John Searle to underline what, in his opinion, is the main difference between the intelligence of man and the one of computational machines.

Let us suppose to lock a person, who does not know even a word of Chinese, in a room, putting at his disposal a large number of cards on which are marked Chinese characters and instructions in Italian. We entrust this person with the task of producing sets of Chinese characters (*output*), using the instructions provided for manipulating the symbols, every time he receives another set of Chinese characters (*input*) from outside. To a Chinese person who asks the questions in writing in his or her own language and receives the answers to them, the person locked in the room seems to understand Chinese perfectly, while in reality that person simply manipulates symbols whose meaning he or she does not understand and assembles them by carefully following the instructions available to him or her. (1)

According to Searle, the fundamental difference between a computational machine and a human being faced with a complex task is the ability *to understand*: a computer is able to perform an assigned task using well-defined instructions stored in its memory; but it is a mechanical, impersonal task performed by a system completely unable to understand the meaning of what it does. A human being, on the contrary, carries out his or her tasks in a conscious way; and this means that he or she is able to understand the links between the various aspects of the problematic situation faced and, in particular, those that link his or her behaviour to what he or she proposes to achieve.

The imaginary experiment proposed by Searle exerts a considerable fascination because it seems to bring a "strong" argument in support of the radical irreducibility of mental faculties to the syntactic operations of computation.

One of the main criticisms of the conclusions drawn from this experiment, which collects much support among the philosophers of mind, is the one put forward by Daniel Dennett. He observes that, in the situation proposed by Searle, if it is true that the human operator dealing with Chinese symbols does not understand the meaning of the Chinese sentences he reads and to which he responds, it is also true that he uses a large number of detailed instructions that guide him step by step in carrying out his task. So, according to Dennett, the operator does not understand the meaning of what he’s doing, but the operator and the instructions he uses constitute a system to which a capacity for understanding must be attributed. (2)

To fully understand the unsustainability of Dennett's objection, let us try to imagine a variant of the experiment proposed by Searle. Let us suppose that the person in the Chinese room is endowed with a prodigious memory, like one of those rare individuals capable of sending an entire telephone book to memory without particular effort. This person would then be able to remember all the instructions contained in the cards we have put at his disposal, perhaps after having read them only once (3). He could therefore provide appropriate answers to questions in Chinese that are addressed to him without the need to consult the cards, relying only on his own mental resources. That person would then be in the condition prescribed by Dennett to manifest the faculty of understanding. (4)

But the ability to elaborate sets of symbols based on pre-existing rules is something that has nothing to do with genuine understanding. Ultimately, that person, although he or she has stored a large number of instructions in his or her memory, would once again limit himself (herself) to their mechanical application, without understanding the meaning of the Chinese texts that he (she) is called upon to elaborate.

Dennett might at this point object that the amount of necessary instructions to give rise to authentic understanding and thus to respond appropriately to questions asked in Chinese is so immense that it is far beyond the reach of any human being. (5)

However why should we accept such an objection? Why is a person of normal intelligence able to learn Chinese and converse in that language, fully understanding the questions asked and answering them appropriately, while the set of instructions needed for a person who doesn't know Chinese to give appropriate answers is so extensive that it cannot be memorized?

Isn't this an admission of the substantial difference between the two situations? Isn't it recognizing authentic understanding, typical of a human being, a clear superiority over the performance of a hybrid system - man-instruction - which, according to Dennett, should give rise to some kind of understanding?

Why then, in order to properly understand the questions, should a person who doesn't know Chinese have the instructions to deal with *any* topic? Why should the lack of instructions to answer some of the questions affect the ability to understand and answer the others correctly?

Let's suppose that the operator locked in the room, subjected to a number of ten questions in Chinese, can only answer the first nine, finding himself in difficulty in front of the last one, concerning a rare butterfly living in the Amazon. The operator, obviously, is unaware that the question refers to butterflies, because he does not know Chinese and does not understand the meaning of what he reads. He simply realizes that he cannot find the appropriate instructions that allow him to transform a group of symbols received into another group of symbols to be sent back to the outside.

However, it doesn't seem to be any rationally acceptable reason why this inability should prevent from the understanding the previous questions, which refer, for example, to philosophy, science or international politics. This consideration makes it legitimate to suspect that the real and only strength of Dennett's thesis is to be sought in its total indemonstrability. Particularly questionable, in this regard, is the claim to use an arbitrary conclusion (i.e. without empirical evidence), deriving from an imaginary situation, as "evidence" to refute an opposing position, equally imaginary.

It has to be said that, pointing out the inadequacy of Dennett's criticism does not mean that the experiment proposed by Searle can be considered a valid argument against the computational thesis of human mind. On the contrary, it can be said that the wide debate to which it has given rise is to be considered artificial and above all sterile, since the possibilities it offers are the result of pure fantasy.

If understanding, and more generally consciousness, are faculties which have arisen and developed from a certain stage of biological evolution, if they are closely intertwined with intelligence in all the problematic situations which we usually have to deal with, there are good reasons to believe that they are capable of conferring some advantage in the management of behaviour: it means, therefore, that their presence is not irrelevant and that, on the contrary, in many circumstances, they must have detectable effects on the results of the activities carried out. We must therefore reject the idea that, in performing a complex task, a person who follows instructions without understanding can perform as well as a person with understanding. In this specific case, the difference between the answers given by a person who knows the Chinese language (and therefore *understands* its meaning) and one who ignores it, however numerous and detailed the syntactical rules available to the latter may be, must always be considered recognizable, at least in principle, by an external observer.

There is no set of instructions for the manipulation of Chinese symbols, however wide it may be, that allows to deal with an unlimited range of subjects, that is, any combination of symbols, making possible a correct conversion into the symbols of another language, while keeping unchanged the meanings they convey.

After all, even the simple translation from one language to another, as long as it is not a simple list of words or elementary phrases, is not just a matter of replacing symbols with equivalent symbols, perhaps keeping in mind some constructs typical of the two languages. In reality - as every translator knows - translation work is much more than simply replacing one word with another, applying the rules of grammar. Behind every language there is a whole world of meanings, experiences, history, characteristic idioms, which are part of a given culture. In addition, the specific context in which a term or phrase is used must be taken into account, as they often take on different nuances or even change meaning depending on the situation or the particular topic in which they are placed. Ultimately, the work of a translator, far from being a mechanical activity to be carried out according to precise rules, which can be explained in an exhaustive way once and for all, is in many ways a work of *interpretation*, which, if on the one hand it is linked to the specific sensitivity of the translator, on the other depends on the characteristics, always changing, of the different contexts to which reference is made.

"Chinese room" experiment, proposed by Searle with the intent to highlight the limits of the computational conception of mind, risks, paradoxically, to bring grist to the mill of the computationalists. Searle, in his arguments, seems to take for granted that there is no detectable difference between the answers given by a person who understands, compared to one who simply follows instructions mechanically. In this way, he shows he doesn't recognize any role to understanding, any positive function: in other words, in his perspective, what distinguishes the action of a human being from the one of a machine is the possession of the ability to understand, but this ability has no effect on the results obtained.

With this, a good part of the relevance of this experiment comes to fall, since the difference between man and machine, from the point of view of intelligence, is reduced to a purely formal question, devoid of any consequence on the empirical field.

NOTES

(The titles of the works cited and the page numbers refer to the Italian edition)

(1) John Searle, "Menti, cervelli programmi", in Douglas R. Hofstadter – Daniel Dennett, *L'io della mente. Fantasie e riflessioni sul sé e l'anima*, Adelphi, Milano, 1985, pp.341-360.

(2) Daniel Dennett, *Coscienza. Che cos'è?,* Rizzoli, Milano, 1993, pag 489. However, it remains to be specified at what level understanding could take place. If it is not the man who understands, who or what actually experiences understanding?

(3) See, for example, the case described by Lurija, concerning a man with a boundless memory, able to remember very long sequences of numbers and meaningless words and to repeat them exactly, even after many years. (Aleksander, Lurija, *Una memoria prodigiosa*, Editori Riuniti, Roma, 1972.

(4) Cf. John Searle, "Menti, cervelli, programmi" cit., pp 346-9.

(5) See Dennett's critique of Jackson's experiment, in Daniel Dennett, *Coscienza. che cos'è,* cit., p. 443 and foll.

[Astro Calisi, *Oltre gli orizzonti del conosciuto...*, pagg 155-159 – English translation by the author]