

A NOTE ON THE CHINESE ROOM¹

Hanoch Ben-Yami

ABSTRACT Searle's Chinese Room was supposed to prove that computers cannot understand: the man in the room, following, like a computer, syntactical rules alone, despite being indistinguishable from a genuine Chinese speaker, does not understand a word. But such a room is impossible: the man would not be able to respond correctly to questions like 'What is the time?', 'What color is this?' etc., even though such an ability is indispensable for a genuine Chinese speaker. Several ways of providing the room with the required ability are considered, and it is concluded that in each case Searle's argument is unsound.

Searle, in his description of the Chinese room, asks us to imagine a man who is locked in a room with a book containing rules in English, specifying which Chinese sentence should be written in response to any Chinese sentence, relying only on the syntax of these sentences. Assisted by these rules, the man can respond so well to questions written in Chinese that are passed into the room that, although he does not understand Chinese, it is impossible for someone outside the room to discern between the answers coming out of the room and those that would have been given by a genuine Chinese speaker. The man inside the room, answering perfectly in Chinese despite his lack of understanding, is analogous to a programmed computer. Searle thus concludes that syntax alone cannot endow a computer with understanding.²

To the best of my knowledge, the objections to Searle's analogy did not oppose the theoretical possibility of his Chinese room.³ By contrast, I shall argue that the Chinese room described above is impossible even in theory, since the man cannot simultaneously be limited to the syntactical level *and* have the competence of a genuine Chinese speaker, and that consequently the analogy is unacceptable. I think many found Searle's claim persuasive, or at least disturbing, because of the apparent theoretical possibility of his Chinese room; showing why it is not possible should, therefore, undermine Searle's position.

Consider the following question written in Chinese and passed into the room: 'What is the time?'. The man in the room will not be able to answer this question as a competent and typical

¹ This is a much-revised version of a paper that first appeared in *Synthese* **95**, 1993, pp. 169-72. The basic idea of the paper is, however, unchanged.

² Searle wrote about the Chinese room in several places. I am relying here on the description given in his 'Minds, Brains, and Programs', (1980; including open peer commentary and author's responses), on Chapter 2 of his *Minds, Brains and Science* (1984), and on his article, 'Is the Brain's Mind a Computer Program?' (1990).

³ Twenty-seven commentaries were appended to Searle's essay in the *Behavioral and Brain Sciences*. For some more recent criticisms see Dennett (1987), and Paul and Patricia Churchland (1990). Dennett and the Churchlands claim that the man is neither able to work fast enough nor to follow a sufficiently complex program in order to imitate successfully a genuine Chinese speaker, and that consequently he does not understand Chinese. It follows that according to them syntax is sufficient for language-understanding: the only obstacles are the speed of processing and the complexity of the program. See also Rapaport (1988, p. 88), who claims 'that being a purely syntactic entity is sufficient for understanding natural language.' I think Searle is right in maintaining the contrary as one of his axioms: 'Syntax by itself is neither constitutive nor sufficient for semantics' (Searle 1990, p. 27).

Chinese speaker would, if he relies only on syntactical rules, as he is supposed to do in Searle's example. If the man is instructed to give an arbitrary answer to such questions, then he will not have the competence to answer them consistently; it would be sheer luck if he gave an answer consistent with his previous one when that question was once again passed into the room half an hour later, say. So in this respect he will not behave as a competent Chinese speaker: his linguistic behavior will not exhibit the *understanding* of what we mean by time. If, on the other hand, he is instructed to answer: 'I don't know; I don't have a watch', then again he will not have the competence of a genuine typical Chinese speaker in this respect. And I cannot see any other way in which such a competence can be imparted to him, given Searle's constraints.

This in itself is no great loss: one can be a fluent Chinese speaker without being able to tell the time—one might simply not have a watch. But obviously, our example is only one of many. We can pass a colored card into the room and ask what color it is, ask the man to say whether a line drawn on a card is straight or curved, play two notes on a piano and ask which one is higher, and so on: to all these he will not be able to answer consistently. The man in the room, relying only on syntactical rules, will equal a blind, deaf, etc., Chinese; i.e., a Chinese bereft of his senses. All such a person is left with are memories, and perhaps the ability to draw some conclusions from them; and such disconnection from the world can indeed be imitated without any understanding, or with next to none.⁴

Examples somewhat similar to mine were given by Dennett (1980, p. 429: 'pass the salt, please') and by Maloney (1987, p. 351: 'prepare tea'). But there is a crucial difference between theirs and mine: while their examples require a nonlinguistic *action*, mine require only an *answer*. I think that, with some reservations irrelevant to our present purpose, Maloney is right in maintaining that 'the ability to affect the environment is not a universally necessary condition of language comprehension in naturally fluent linguistic agents' (1987, p. 352). And my examples show that the man in the room cannot be a 'fluent linguistic agent' even if such an agent is not required to affect the environment in any direct manner, apart from the production of speech. (See also the related point on the Robot Reply below.)

There are only two alternatives open to us, I think, if we want to endow the man in the room with the competence of a genuine Chinese speaker. The first is to include among the rules (i.e., the program) instructions of this kind: 'If you are given the sentence ... (here the Chinese question corresponding to the English 'What is the time?') should be written) *look at a clock* and write a sentence according to these rules: ...'. In this alternative, the man responds correctly by himself. But in this case, the man's ability to answer the questions relies on his *understanding of the concepts* of time, color (in my second example), etc. If he is analogous to the programmed computer, then we have to conclude that the computer has *understanding*.

The man might of course still not understand Chinese: he might fail to see the relation between the time his watch shows and the sentences he writes. But this is obviously beside the point. What is at issue is whether, in order to reply as a genuine Chinese speaker would, the man must understand the *concepts*—not the *words*—contained in the Chinese sentences passed to him. Similarly, suppose we had *two* men in the room, one translating, according to mere syntactical

⁴ I now think this is insufficient as it stands: this ability does generally express intelligence and understanding. Accordingly, my argument in this paper is not entirely sufficient to show the inadequacy in Searle's thought experiment. I think that the argument I developed in (Ben-Yami 2005) against Ned Block's partly similar thought experiment (Block 1980) explains more clearly what is wrong with this residue: in this case the man in the room reproduces replies given to him by someone else (the programmers), and *this* ability does not express intelligence or understanding. Similarly, a computer that is limited to reproducing pre-stored data does not exhibit understanding.

rules, the Chinese questions passed into the room into Hungarian, without understanding a word in neither language, and the other, a fluent Hungarian speaker who speaks no other language, replying to them in Hungarian in sentences that are then translated back into Chinese by the first. No one in the room would then understand Chinese, but that ability to reply to questions in Chinese would rely on the understanding of the concepts involved (by the Hungarian). The fact that no one then understands Chinese would not show that a computer that replies as a genuine Chinese speaker would does not have understanding.

The second possible way to enable the man in the room to answer the question successfully is not to make him respond by himself, but to instruct him to pass the question on, possibly after some syntactical processing, to someone or something that will supply him with the correct answer (for instance, a clock can stamp the time, in Chinese characters, on a card); and the man in the room will only have to pass it on as the output, maybe after some additional syntactical operations. In this case, the man in the room remains on the syntactical level, while whatever supplies him with the answer has to refer to a clock. In such a case, the existence of understanding depends on what we include in our view. If we consider only the man in the room, then his operations do not depend on any understanding of the content of the questions passed to him; but then neither is he equivalent to a genuine Chinese speaker nor, therefore, to a computer which is equivalent to such a speaker; hence, we cannot conclude from the lack of understanding in his operations that the computer lacks understanding. If, on the other hand, we look at the man in the room combined with what is passing the answers to him, then we do have in this union the equivalence to a genuine Chinese speaker. But since an important part of the processing is not done by the man in the room, we again cannot conclude from his lack of understanding that a computer will lack understanding.

Accordingly, on both alternative modifications of the Chinese room, Searle's argument does not work.

The objection most similar to mine I could find in the literature is the one Searle dubbed 'the Robot Reply' (Searle 1980, p. 420). This is his formulation of the objection:

Computers would have semantics and not just syntax [i.e., what they say or write would have meaning and therefore they could be said to have understanding] if their inputs and outputs were put in appropriate causal relation to the rest of the world. Imagine that we put the computer into a robot, attached television cameras to the robot's head, installed transducers connecting the television messages to the computer and had the computer output operate the robot's arms and legs. Then the whole system would have a semantics. (Searle 1990, p. 30)

But the difference between this objection and mine is essential. The Robot Reply takes the contact with the world as a *contingent addition* to the functioning Chinese room, one without which it may still have a linguistic competence similar to that of a genuine Chinese speaker. By contrast, I think such an interaction with the world is *necessary* in order to make the room function at all: without it, it could not have the linguistic capacities of a genuine speaker. Searle convincingly rebuts the Robot Reply in (Searle 1980, p. 420).

In conclusion, we can say that Searle was perhaps right in arguing that mere syntactical rules do not suffice for understanding, but also that he was definitely wrong in assuming that a computer, successfully imitating our linguistic abilities, can be limited solely to syntactical rules (the Chinese room). Such a computer will have to interact with things in the world, be it a watch fixed inside it or perceptual information available via various instruments, in order to achieve the required competence. 'Computer programs are formal (syntactic)' (Searle 1990, p. 27), that's for sure; but the computers in which they are realized make contact with the surrounding world in various ways, essential for their functioning. Accordingly, the formality of computer programs does not prove

that the computers themselves lack understanding.

*Philosophy Department
Central European University
Nádor u. 9, H-1051 Budapest
Hungary*

REFERENCES

- Ben-Yami, H. 2005 'Behaviorism and Psychologism: Why Block's Argument Against Behaviorism is Unsound', *Philosophical Psychology* **18**, 179-86.
- Block, N. 1981 'Psychologism and Behaviorism', *Philosophical Review* **90**, 5-43.
- Churchland, P.M. & Churchland P.S. 1990 'Could a Machine Think?', *Scientific American* **262**, 32-7.
- Dennett D. 1980 'The Milk of Human Intentionality', *The Behavioral and Brain Sciences* **3**, 428-30.
- Dennett D. 1987 'Fast Thinking', in his *The Intentional Stance*, Cambridge, Mass.: MIT Press, 323-37.
- Maloney, J.C. 1987 'The Right Stuff', *Synthese* **70**, 349-72.
- Rapaport, W.J. 1988 'Syntactic Semantics: Foundations of Computational Natural-Language Understanding', in J.H. Fetzer (ed.) *Aspects of Artificial Intelligence*, Dordrecht: Kluwer Academic Publishers, 81-131.
- Searle, J. 1980 'Minds, Brains, and Programs', *The Behavioral and Brain Sciences* **3**, 417-57 (including open peer commentary and author's responses).
- Searle, J. 1984 *Minds, Brains and Science*, Cambridge, Mass.: Harvard University Press.
- Searle, J. 1990 'Is the Brain's Mind a Computer Program?', *Scientific American* **262**, 25-31.