

CAN MACHINES THINK?

Paulo R. Margutti Pinto

Dept of Philosophy

Universidade Federal de Minas Gerais

[Published in *Ciência e Cultura*. Journal of the Brazilian Association for the Advancement of Science, São Paulo - SP, v. 50, n. 2/3, p. 149-152, 1998]

ABSTRACT:

The approach that explains the mind on the basis of mental processes is contrasted with the approach that explains the mind without appealing to such processes. The contrast is made on the basis of Turing's "Imitation Game", in which a human judge communicates with two screens. One of them is controlled by a human being and the other by a computer. Without knowing who is who, the judge is supposed to decide who is hiding behind each screen. In case the judge fails, the computer wins the game. Although the adherents to the thesis that machines cannot think have good arguments for their case, their opponents always seem to have a good and disturbing reply to such arguments. In addition, the only feature which distinguishes human beings from computers, namely the ability to experience the mystical, paradoxically cannot be used as a criterion in Turing's Game.

The problem

In a seminal paper, entitled 'The Imitation Game', A. N. Turing suggests an experiment that would allow us to decide whether machines can think or not. Suppose a human being, a computer, and a judge. Both the human being and the computer would be asked questions by the judge and would print their respective answers on screen. The judge would not be able to see who is behind a particular screen. In this way, he or she would not be able to see who is answering. The experiment is described as a 'game' because the computer would win in case the judge either wrongly identifies it with the human being or is unable to identify any of the participants as a machine. According to Turing, the experiment is decisive: if the computer wins, we would have to reckon that machines can think. According to other thinkers, who tend to reaffirm the theses connected with traditional philosophy, the experiment is not decisive. Turing's claim seems to be far too exotic to be true. However, if we go deeper into this debate, we shall see that his claim may be made more palatable. What is more, the

discussion suggests that the real difference between human beings and machines may be situated somewhere beyond rational discourse.

Philosophy of Mind with a mind

The philosophers that challenge the decisiveness of Turing's Game claim that thinking involves mental processes which do not occur within the computer. We think by means of concepts, which are abstract, universal and immaterial. We may say so because we access our thinking through introspection, that is to say, through a process of internal observation, in which the thinking subject observes himself or herself. This makes our thinking purposeful and versatile, in the sense that it is directed towards an end and may be applied to a great number of different situations. From this standpoint, thinking is an internal process that cannot be confused with the results of the process itself.

All the above claims converge into the main claim that thinking is irreducible to explanations which appeal either to matter or to mechanical processes. And since machines can be explained in this way, it follows that they cannot think. The only beings that can think are human beings.

In order to prove that mind is irreducible either to matter or to physical processes, and consequently that machines cannot think, several arguments have been advanced. Two of them are worth mentioning here: the argument concerning Gödel's Theorem and the argument concerning the Chinese Room.

Consider first the argument concerning Gödel's Theorem. In 1931, Kurt Gödel, a very young mathematician aged 26, obtained an astonishing result concerning formal systems. He analyzed the formal system expounded in "Principia Mathematica", one of the chief works in the domains of logic and mathematics in our century. The authors, B. Russell and A. North Whitehead, succeeded in constructing a formal system which combines consistently Symbolic Logic and Arithmetic. They assumed that every true mathematical sentence would be demonstrable within the system. Now Gödel showed that any formal system of this kind would allow the construction of a very peculiar sentence, which would be at the same time true and not demonstrable. This is a very technical result and can only be explained roughly.

In his Theorem, Gödel establishes an extremely ingenious numbering system, in which every sentence corresponds to one and only one number (call it its Gödel Number). Thus, given any sentence belonging to the system in "Principia", we are able to give its Gödel Number, and conversely, given

any Gödel Number, we are able to give the corresponding sentence in “Principia”. The numbering system allows the construction of a particular sentence ‘G’, of which the Gödel Number is n and which affirms ‘there is no proof for the sentence with Gödel Number n ’. Thus, ‘G’ affirms of itself that it has no proof. This has an awkward consequence: if there is a proof for ‘G’, then the system would be contradictory, because it would be proving a sentence ‘G’ affirming precisely that ‘G’ has no proof. In order to avoid contradiction, we must accept that ‘G’ cannot be proved, or, in other words, that ‘G’ is true. In this way, Gödel succeeds in constructing a sentence that is simultaneously true and not demonstrable within the system. This means that there is at least one true sentence that cannot be proved by the formal system in “Principia Mathematica”. Of course, the original Theorem is much more sophisticated and rigorously formulated than it is in the above exposition. But we are here interested in Gödel’s result, and not in the way he obtained it.

J. R. Lukas thinks that Gödel’s Theorem provides a means to discriminate between the computer and the human being in the Imitation Game. Suppose the judge asks some question concerning the paradoxical sentence ‘G’. The human being would easily recognize it as a true sentence, although not demonstrable, and would answer accordingly. The computer would be in a very different situation. As a machine, it has been designed to deal only with sentences which are at the same time demonstrable and true. Therefore, it would be incapable of calculating a sentence which is simultaneously true and not demonstrable. I would not be able to recognize ‘G’ as true, although not demonstrable. Thus, the answers concerning ‘G’ would be different and the judge would then be able to decide, without mistake, between the human being and the machine. This is so because only the mind, which is self-conscious, would be able to affirm its own consistency, recognizing ‘G’ as a true but not provable sentence. The computer is unable to do so. This is the argument concerning Gödel’s Theorem, in favor of the irreducibility of the mind to a machine.

Consider now the argument concerning the Chinese Room. Suppose a person is locked inside a room with some material written in Chinese. This person speaks English, but does not know anything about the Chinese language. In the room there is also a set of rules in English which allows the person to correlate certain groups of Chinese symbols with other groups of Chinese symbols, on the basis of their external form alone. The situation may be structured in a way such that a Chinese speaking person might introduce a question in Chinese into the room (through a small window, for instance); the corresponding symbols might be processed by the person inside the room, according to the set of English rules; this would yield a group of Chinese symbols which would be returned (through the same window) to the Chinese speaking person and would be recognized as a perfectly sound answer in

Chinese. Of course, the person inside the room might receive questions in English and answer to them in good English as well. But in the first case, the person inside the room would be working just like a computer: he or she would be performing mere calculations with elements defined by their external form. There would be no understanding at all. In the latter case, the person inside the room would be working just like our mind: he or she would be performing more than mere calculations on the basis of the form of the symbols. The understanding of the symbols would be involved. This ‘thought experiment’ has been designed by the philosopher John Searle and its purpose is to show that mere calculations with symbols do not explain thinking. Something else is required in order to explain the understanding which accompanies such calculations.

The above arguments converge into the notion that language is a means for expressing our universal and immaterial thoughts. There exists, prior to any language, the language of thought, which consists of connections between abstract concepts. We may call this thought language ‘Mentalese’. So, computers do not think because, although they make mechanical calculations with symbols, they are not able to understand Mentalese.

Philosophy of Mind without a mind

The philosophers who admit the decisiveness of Turing’s Game claim that thinking may be explained without the appeal to mental processes. They consider that the notion of a ‘concept’, understood as an abstract, universal and immaterial entity, is a dispensable fiction. They are also extremely suspicious of ‘introspection’ as a privileged access to our thinking through the “eyes” of the mind. For them, the notion of an inner process of self observation is also dispensable. It is true that our thinking is purposeful and versatile, in the sense that it is directed towards an end and may be applied to a great number of different situations. But since the appeal to internal processes is dispensable, our thinking may be evaluated by the results of the processes themselves.

It is clear that the above claims converge into the main thesis that thinking is reducible to explanations which appeal either to matter or to mechanical processes. And since machines can be explained in this way, it follows that they can think. Human beings are no longer the only beings that can do this job.

These ideas are originated from the works of contemporary philosophers, such as Ryle, Wittgenstein, Sellars, Quine and Davidson. As an illustration, consider some ideas of Wittgenstein and Quine on the subject.

Wittgenstein attempts to show that “mental processes” are not required in the explanation of how language works. He argues that the appeal to such “processes” is the result of the bewitchment exerted by language upon our minds. For example, when someone is walking and says “I walk”, there is a visible physical process which accompanies the sentence. By analogy, we tend to affirm that, whenever someone says “I think”, there will be an invisible “mental process” accompanying the sentence. But the analogy is mistaken. It depends on some kind of fascination that results from observing only one or two of the relevant aspects of the functioning of the word. The meanings of our words are yielded by the multifarious uses we make of them. And we use the word ‘think’ in so many ways that we could hardly find something as a common, stable and prior “meaning” for the word. This makes dispensable the appeal to the “mental process” that would be required in order “to grasp” the “meaning” of the sentence “I think”. If we pay attention to the diversity of aspects involved in the functioning of language, we will be free from the bewitchment generated by assuming a unilateral standpoint.

In an analogous spirit, Quine rejects the notion that there is something like “the meaning” in itself. We learn how to speak a language by observing the behavior of our interlocutors and by making hypotheses about the meanings of their utterances. Mistaken hypotheses may be corrected by observing ulterior behavior of the interlocutors concerning the same utterances. The problem is that different hypotheses may be made and many of them may be compatible with all the facts. For example, suppose an ethnologist that is trying to translate the language of a tribe into English. One of the natives sees a rabbit and exclaims ‘gavagai!’. The ethnologist might take it to mean ‘there is a rabbit’, or ‘this is rabbit stuff’, or ‘rabbithood is instantiated over there’, or ‘a stage in the history of a rabbit is over there’, etc. Each hypothesis would yield a particular translation manual, which would involve attributing a different meaning to the expression ‘gavagai’. Quine claims that many of the translation manuals would fit all the facts concerning the behavior of the natives whenever they use the expression ‘gavagai’. If this is correct, then we would be led to the inevitable conclusion that we cannot access “the meaning” of ‘gavagai’ in itself because it does not exist. The world of Mentalese is a fiction. So, we cannot say that the translation manual adopted is good or bad; we cannot say that some translation from, say, French to English, is better than another. The problem does not make sense, because there are no fixed meanings.

Quine also argues that notions like ‘analyticity’, ‘synonymy’, ‘necessity’, ‘essence’, and ‘intensionality’ are ill-defined and useless. They should be replaced by notions like ‘true’, ‘false’, ‘not’, ‘or’, ‘and’, ‘if-then’, etc. The latter are in better shape to deal with reality.

The ideas of Wittgenstein and Quine tend to encourage the philosophers which adhere to the claim that machines can think. In fact, the two approaches suggest that we should look for alternative ways of explaining thinking, and that this perhaps may be done by appealing only to the more manageable notions above mentioned. Such notions are the ones computers are able to deal with. Similar claims may be made concerning the works of Ryle, Sellars, Davidson and others.

On the basis of such ideas, the adherents to the claim that machines can think argue that thinking does not necessarily involve mental processes. We do not know what it is to characterize thinking as a mental process. We may abandon the mysterious process of understanding abstract concepts and try to infer the thinking from the results obtained. As opposed to the inaccessible private mental processes, the latter are publicly observable. We may recognize the presence of thinking by merely observing the corresponding behavior. In this sense, the fact that a computer plays good chess would be strong enough a criterion to decide whether the computer is thinking or not. And some day a computer will possibly win the Imitation Game. That is why Turing’s Game is decisive.

As a reply to the argument that a computer could never win the game in virtue of the limitations imposed by Gödel’s Theorem, it may be claimed that an adequately prepared machine would be able to deal with sentence ‘G’. The software would be designed in a way such that the computer would recognize a Gödelian sentence and adopt the adequate procedures in order to avoid the appearance of being at a loss. This would impede the judge to distinguish the human being from the machine, at least as far as Gödel’s Theorem is concerned. In addition, it may be also claimed that the appeal to the thesis that only the mind is capable of affirming its consistency is too vague. The thesis is based on notoriously ill-defined notions, such as ‘self-consciousness’.

As a reply to the Chinese Room argument, according to which thinking involves understanding, it may be claimed that ‘understanding’ is an obscure notion. We do not need meanings or interpretations in order to explain thinking. The publicly observable interactions between the thinking subject and the world are more than enough. Thus, we are allowed to say that a person understands what the word ‘car’ means because he or she reacts properly in all situations involving cars. But the same is valid for computers as well. Machines and thinking are not incompatible at all.

The above arguments converge into the notion that language is not a means for expressing our universal and immaterial thoughts. There is no such thing as “Mentalese”. The appeal to “mental

processes” is dispensable. We may say that, although the computer merely makes mechanical calculations with symbols, it in fact thinks, because some day in the future the judge in the Imitation Game will not be able to distinguish the human being from it on the basis solely of its behavior.

True, the line of argumentation adopted reveals that the meaning of ‘thinking’ itself has been altered. Whereas the adherents to the approach of a philosophy of mind with a mind conceive of ‘thinking’ as inseparable from mental processes such as understanding, the adherents to the approach of a philosophy of mind without a mind conceive of ‘thinking’ merely by its publicly observable results. This is, of course, to give a new meaning to the word ‘thinking’. But the main point is that the new meaning is a consequence of a new way of seeing things, in which the appeal to the problematic “mental processes” is abandoned. Whether this new approach is correct or not remains to be seen. One thing, however, is for sure: it seems to be much more promising than the traditional ones.

Minds, machines and mysticism

The adherents to the claim that mind is irreducible to matter or to mechanical processes may still appeal to a last argument. They may say that human beings are capable of some sort of experience that is beyond the reach of machines. This is the mystical experience, which usually includes the logically contradictory feeling that everything is equal to everything, that all oppositions fuse into a higher totality. William James characterizes such an experience by four marks: i) ineffability (it defies expression: no adequate report of its contents can be given in words); ii) noetic quality (although very similar to states of feeling, the mystical experience seems also to be a state of knowledge); iii) transiency (it cannot be sustained for a long time); iv) passivity (it is also the experience of being grasped and held by a superior power). If this is true, we may say that there is in fact a difference between human beings and machines.

But this is so peculiar and special a difference that the adherents to the thesis that thinking is explicable by mechanical processes may still have a reply. In fact, we know that the mystical experience - if any - is in itself inexpressible. If, on the one hand, we ignore this feature and try to talk about it, then the discourse about the mystical experience may be included in the scope of the Imitation Game. In this case, the computer may be programmed in a way such that its discourse about the mystical is indistinguishable from any human being’s discourse on the same subject. The computer would be able to win the game. If, on the other hand, we seriously recognize that the mystical

experience is inexpressible, then the discourse about the mystical experience would be excluded from the scope of the Imitation Game. In this case, neither the human being nor the machine would be able to say the unsayable. The ability to have mystical experiences, although it may constitute a distinguishing feature of human beings, paradoxically lies outside the domain of discourse and cannot be detected by the Imitation Game.

REFERENCES

1. Turing AN 1964. The Imitation Game, *In Minds and Machines*, Anderson AR org. Prentice Hall, Englewood Cliffs, N.J.
2. Nagel E, Newman JR 1973. *Gödel's Proof*. New York Un. Press, N.Y.
3. Lucas JR 1964. Minds, Machines and Gödel, *In Minds and Machines*, Anderson AR org. Prentice Hall, Englewood Cliffs, N.J.
4. Searle J 1981 Minds, Brains and Programs, *In The Mind's I*, Hofstadter, D, Dennett, D orgs. Basic Books, N. Y.
5. Wittgenstein L 1975 *The Blue and Brown Books*, Basil Blackwell, Oxford.
6. Quine WvO 1960 *Word and Object*, MIT Press, Cambridge, Mass.
7. James W 1985 *The Varieties of Religious Experience*, Penguin Books, Harmondsworth, Middlessex, England.