Chinese room argument in the era of non-symbolic AI

Simone Reale, Politecnico di Milano

Abstract

Today artificial intelligence represents an inescapable presence, from chatbots to virtual assistants, it has established itself as the dominant feature of the 2020s. More than ever it is necessary to ask one of the questions that stimulated the curiosity of the founding fathers of the discipline: "Can machines think?". One of the most important names related to this question is that of John Searle, his thought experiment commonly called "Chinese Room Argument" still represents a milestone in the academic debate on the topic. This paper aims to verify the validity of the aforementioned thought experiment in relation to the current state of technology, in particular, it aims to verify the idea that the semantic content of mental processes cannot be replicated by formal systems, it will try to demonstrate how in some conditions semantics is an emergent property. The analysis will be based in particular on the use of computational paradigms commonly labeled as non-symbolic with examples taken from models specialized in image classification and natural language processing.

Keywords

Artificial Intelligence, Searle's Chinese Room Argument, Artificial Neural Networks, Cognitive Science

1. Introduction: why should we care about thinking machines?

In today's world, where artificial intelligence is starting to claim a prominent role within everyone's private and public lives, it is of paramount importance to discuss its limits and prospects.
It is necessary to revisit the fundamental concepts in this field in order to determine whether they have stood the test of time, just like in any field that experiences a sudden technological acceleration.
In the 50s, Computer Science studies were just starting to develop in a structured manner, and the challenging question "Can machines think?" was beginning to make its way into universities and research centres. The environment was characterized by optimism as a direct consequence of the rapidity with which progress was being made in the field, an optimism accurately captured in the words of Alan Turing: "I believe that at the end of the century the use of words and general educated opinion will have changed so much that one will be able to speak of thinking machines without expecting to be contradicted.".

In this context, Turing pioneered the homonymous test as a criterion for establishing whether a machine was capable of exhibiting behaviour deemed "intelligent".

After more than fifty years, the debate is still ongoing, and it hasn't lost its primary relevance. On the contrary, technology's latest improvements have significantly highlighted how the issue affects us even more profoundly.

AI is making its way into fields where traditionally there has never been room for anything other than human intellect: a future in which machines will be able, for example, to

provide direct support to the legal system or private and public administration does not seem too distant to us. In cases like the aforementioned, it is evident how the question “Can machines think?” and the corollary “Do we really want machines that think?” assume renewed importance.

One of the seminal studies on the issue is the thought experiment conceived by John Searle, commonly called the Chinese Room Argument. Its influence is corroborated by the considerable number of citations it still receives to this day.

The Turing test and Searle’s conjecture are often mentioned together, even though they paint two different visions of the concept of intelligence. The former assumes that it is equated to behaving intelligently (the machine tries to deceive the interviewer and appear humanly intelligent), and in some way, this approach has something in common with the issue of mutual understanding between interlocutors within the theory of language games by Ludwig Wittgenstein. For Turing, the question “Can machines think?” seems to lose its meaning, while for Wittgenstein the same happens to the question “Was there actually understanding?”: conclusions are consequently drawn from the sole observation of external behaviour, but what do these questions have in common?

Both have to do with the seemingly unfathomable nature of thought processes.

Searle, on the other hand, tries to disprove the validity of the Turing Test by demonstrating how outward behaviour is not an effective index of real cognition. He proposes the theory of biological naturalism which affirms the “reality” of mental processes but also their irreducibility to cerebral states (particular sequences of neuronal activations), an alternative to the classical stances of materialism and dualism on the topic.

This “reality” breaks the veil of mysticism that hovers over mental processes by treating them as biological phenomena like others, allowing a more direct treatment of the problem, which is why Searle and his argumentations will be the starting point of our analysis.
His conclusions can be summarized in three related propositions:
"1: Syntax is not sufficient for semantics;
2: Programs are formal;
3: Minds have content.” [1]
The ultimate goal of this paper is to challenge the first claim, using state-of-the-art technology to demonstrate how semantics can be considered an emergent property of syntax.

2. The Chinese Room Argument

In his “Chinese Room” argument Searle tries to refute the possibility of creating Strong AI based solely on formal and syntactic abilities, namely a machine whose intellectual capability is functionally equal to a human’s. In his words: “... according to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states. In strong AI, because the programmed computer has cognitive states, the programs are not mere tools that enable us to test psychological explanations; rather, the programs are themselves the explanations.”[2]

To summarize his argument let’s imagine that an individual is locked up in a room. This person is a native English speaker and does not understand the Chinese language, neither written nor spoken. In the room, this individual finds a sheet full of Chinese ideograms and a second sheet, always rigorously written in Chinese, with a series of questions. The man is therefore faced with two series of symbols that have no meaning for him, but in the room there is also a book with a series of rules written in English that explain how to match the symbols of the first sheet with those of the second sheet.

Suppose the first sheet is a story written in Chinese and the second a series of questions about the story. The man begins to produce a

response output, following the instructions given to him. In this example, the instructions represent the computer program. The answers produced are formally correct because the man has followed the instructions given to him to the letter together with the ideograms. Despite this, he didn't understand anything he received, or what he replied, and, of course, he had not learned anything. If there were any external observer to the experiment, however, let's also say a native Chinese speaker, he could think that the man had a good knowledge of his language.

According to Searle, in the same way that a man mechanically executes the order without understanding Chinese, the computer executes the program written in the programming language which is, in some sense, his native language, but essentially manipulates symbols of which he does not know the meaning.

So the fulcrum of his arguments revolves around the observation of the purely formal and syntactic nature of programs in contrast to the semanticity of mental processes, semanticity derived from intentionality that is "the power of minds and mental states to be about, to represent, or to stand for, things, properties and states of affairs " [3], a characteristic which according to him belongs only to the realm of the biological.

Ludwig Wittgenstein can be seen as a forerunner of some of Searle's arguments and his language-focused point of view can help clarify the problem further. A thought experiment similar to Searle's was proposed by him years earlier: "Suppose I train the apprentices of wallpaper manufacturers so that they can produce perfect proofs of the most complicated theorems in higher mathematics [and use these as wallpaper patterns], in fact so that if I say to one of them 'Prove so-and-so' (where so-and-so is a mathematical proposition), he can always do it. And suppose that they are so unintelligent that they cannot make the simplest practical calculations. They can't figure out if one plum costs so-and-so, how much do six plums cost, or what change you should get from a shilling for a twopenny bar of chocolate.-Would you say that they had learnt mathematics or not?... ".
It is evident how for both of them what distinguishes a sensible sentence from a meaningless one are the thoughts that it provokes in the subject, the semantic representation that is created internally by the subject.

3. Does technology matter?

John Searle formulates his arguments in a period in which artificial intelligence is still linked almost exclusively to classical computation.
He strongly supports the independence of his conclusions from the particular technology used (both hardware and software) and thus proposes the absolute impossibility of obtaining real cognition starting from simple symbol manipulation. But is he right? Are the ultimate conclusions of the Chinese Room Argument independent of the technology used?
It is evident that the irreducibility of the mind to the brain clashes with the transparency of the path taken by an input to become an output in a system based on the classical notions inspired by Turing. The analogy between these systems and the human mind does not seem to hold, to find something more promising we must therefore look at systems in which the transparency previously mentioned is not present and the link established between input and output is not obvious.

4. Artificial Neural Networks

Since artificial neural networks are directly inspired by the biology of animal brains' ones, there seems to be an encouraging possibility, but what is their structure, and why is it important?
ANNs are composed of several layers of artificial neurons that act as basic units of computation. Each artificial neuron is mutually connected to various others and it has a distinctive associated weight and threshold.
In case of a node's output exceeding the specified threshold value, that node is

triggered, sending data to the next layer of the network. Otherwise, no information is transmitted to the next layer of the network and data computed by the individual layers are then combined to obtain the final output.

An objection could concern the distance between the state-of-the-art models and the starting biological model, ANNs in fact have gradually abandoned some assumptions which have been at the basis of the development of the technology and which derive directly from the observation of brain processes. Artificial neurons, for example, have increasingly distanced themselves from their natural counterparts in their basic functioning. In any case, the most common opinion is that the greater or lesser biological plausibility is not strictly linked to the potential of the model to produce mental processes: " [...] it is premature to draw firm conclusions based upon biological plausibility, given how little we understand about the relationship between neural, computational, and cognitive levels of description." [4]

One could also reasonably argue that ANNs are formally expressive at the same level as classical computational models: it is possible to implement a neural network in a classical system and vice versa, and this idea seems to downsize the scope of the change they bring.

The real power of the ANNs, however, lies in the notion of symbol they suggest which is very different from what Searle had in mind when he wrote his arguments. Investigating further on this concept can help clarify how revolutionary the contribution of this approach is.

In fact, in the ANNs, information is not all concentrated in easily traceable symbols, but rather it is scattered throughout the network in the form of weights in the connections between one neuron and another, activation functions, and all the meta-parameters that manage the functioning of the system. This has even caused some scholars to consider this computational paradigm non-symbolic, at least in the classical sense, a more robust notion of symbol has become necessary: "Many discussions of the symbolic/non-symbolic dichotomy employ a more robust notion of "symbol". On the more robust approach, a symbol is the sort of thing that represents a subject matter. Thus, something is a symbol only if it has semantic or representational properties. If we employ this more robust notion of symbol, then the symbolic/non-symbolic distinction cross-cuts the distinction between Turing-style computation and neural network computation.

A Turing machine need not employ symbols in the more robust sense. As far as the Turing formalism goes, symbols manipulated during Turing computation need not have representational properties [5].

Conversely, a neural network can manipulate symbols with representational properties. Indeed, an analog neural network can manipulate symbols that have a combinatorial syntax and semantics." [6]

Now it is easy to see how the previously mentioned "transparency" has disappeared, the connection between input and output and the transformation process in the middle are no longer clearly visible, replicating, in a sense, the relationship between mental and brain processes.

An example that could clarify what the representational property of the symbols used by ANNs consists of can be offered by the functioning of a type of network, called convolutional, used specifically in image recognition.

Typically in the ANNs, the output of a neuron depends directly on all the input to the network, in the case of convolutional networks instead, the output depends only on a limited region of the input image, this input region is called the receptive field. This difference means that convolutional networks can operate on different levels of granularity, but how does this relate to the property described previously?

To understand this it is necessary to observe the intermediate results in the network: let's take, for example, the image of a cat as input; gradually, we proceed towards the final result of the analysis within the network and the intermediate images generated represent progressively more complex characteristics: starting from the general shapes we move on to the shape of an ear, an eye or a nose, all of this

is encoded in the weights, activation functions, and all the necessary parameters.
Raw data is therefore supplied to the network as input and to obtain a result it needs to recognize patterns and create its complex, high-level representations: essentially, it needs to add a semantic characterization.
After having clarified the innovative scope of models based on artificial neural networks, it is natural to re-discuss the Chinese Room Argument and how it can be treated in the light of what has been said.
Firstly, it is necessary to talk about the state of the art of natural language processing.
The modern approach related to the natural language processing starts by considering single terms as atomic symbols in a discrete space made up of all the words considered. The next step consists in transforming this space by reducing its dimensionality and making it continuous.
In doing so, arbitrariness is introduced: the mapped position of individual words and the mutual distance can represent the embryo of a semantic representation.
To better contextualize, we can take, for example, the "Word2Vec" model developed in Google research centres [7].
In this model, the space learned by the system after a training period presents surprising regularities: for example, the vector difference between the words denoting the name of a state and those denoting its capital and the distance between nouns in the feminine and masculine ones are constant.
As far as the position in absolute terms is concerned, however, it is interesting to see how words with similar meanings are grouped in clusters: since the model is hyper-dimensional a word can belong to different clusters of meaning along different axes, for instance, the word "Paris" can be in direct proximity to others denoting further capitals along one axis, and near others describing its points of interest on another.
Even more interesting to observe is how the typical vector sum and difference operations can, in this space, work on meanings obtaining results that seem to be the result of human intuition.
To better understand also in this cases some examples are needed:
"Einstein"-"Scientist" + "Painter" ~ = "Picasso" or "His"-"He" + "She" ~ = "Her".

5. Conclusion

Chomskian linguistic theory can help to characterize the nature of the aforementioned space: he explains how it is possible to divide human language into two classes: extensional and intensional [8].
The first is the external language used to communicate (what we commonly call idiom) and has a well established grammar; the intensional language, on the contrary, is devoid of grammar, it is the secret and individual language that operates in depth and defines the space that enclose the semantic representation of the concepts we talk about.
An analogy between the previously defined space generated by ANNs and the one determined by intensional language does not seem absurd, if we accept it then it is possible to look at the Chinese Room Argument from another perspective.
We could in fact argue that if the man in the "Chinese" room, instead of using a classical computational paradigm, used that of artificial neural networks, he could, with training, intuit the structure of the intensional language and therefore obtain a semantic representation of what he reads.
All this is in direct contrast with what Searle states, given that artificial neural networks, as previously mentioned, can be implemented on classical computational systems and are in all respects computer programs. In this light, we can describe the semantics derived from the ANNs as an emergent property of syntax.

References:

[1] Searle, J. (1994), "The Mystery of Consciousness", (London: Granta Books)

[2] Searle, J. (1980), "Minds, Brains, and Programs." Behavioral and Brain Sciences 3, 417-424.

[3] https://plato.stanford.edu/entries/chinese-room (last visited January 2022)

[4] Gallistel, C.R. and King, A., (2009), "Memory and the Computational Brain", Malden: Wiley-Blackwell.

[5] Chalmers, D., (2011), "A Computational Foundation for the Study of Cognition", The Journal of Cognitive Science, 12: 323–357

[6] Horgan, T. and J. Tienson, (1996), "Connectionism and the Philosophy of Psychology", Cambridge, MA: MIT Press

[7] Mikolov, Chen, Corrado, Dean , (2013), "Efficient Estimation of Word Representations in Vector Space"

[8] https://www.rep.routledge.com/articles/biographical/chomsky-noam-1928/v-1/sections/the-aims-and-principles-of-linguistic-theory

(last visited January 2022)