Vilnius
University
Press

Contents lists available at Vilnius University Press

Problemos
2019, vol. 96, pp. 121–133

ISSN 1392-1126 eISSN 2424-6158
DOI: https://doi.org/10.15388/Problemos.96.10

Pažinimo ir sąmonės filosofija / Philosophy of Mind

A Fair Version of the Chinese Room

Hasan Çağatay
Social Sciences University of Ankara
E-mail: hasan.cagatay@asbu.edu.tr

Abstract. By the Chinese room thought experiment, John Searle (1980) advocates the thesis that it is impossible for computers to think in the same way that human beings do. This article intends firstly to show that the Chinese room does not justify or even test this thesis and secondly to describe exactly how the person in the Chinese room can learn Chinese. Regarding this learning process, Searle ignores the relevance of an individual's pattern recognition capacity for understanding. To counter Searle's claim, this paper, via examining a series of thought experiments inspired by the Chinese room, aims to underline the importance of pattern recognition for understanding to emerge.

Keywords: Artificial intelligence, Chinese room, Turing test, understanding, pattern recognition

Teisingesnė „kinų kambario“ versija

Santrauka. Naudodamasis „kinų kambario“ mintiniu eksperimentu, Johnas Searle'as (1980) gina teiginį, jog kompiuteriai negali mąstyti taip, kaip mąsto žmonės. Šiame straipsnyje pirmiausia ketinama parodyti, kad „kinų kambario“ eksperimentas ne tik kad nepagrindžia, bet net ir neišbando šios tezės, o, antra, paaiškinama, kaip kinų kambaryje sėdintis žmogus gali išmokti kinų kalbos. Kalbėdamas apie šį mokymosi procesą Searle'as ignoruoja tai, kokią svarbą supratimui turi asmens gebėjimas atpažinti struktūras. Nesutikdami su Searle'o teze, šiame straipsnyje nagrinėjame keletą kitų, kinų kambario įkvėptų, mintinių eksperimentų ir pabrėžiame struktūros atpažinimo svarbą supratimui atsirasti.

Pagrindiniai žodžiai: dirbtinis intelektas, kinų kambarys, Turingo testas, supratimas, struktūros atpažinimas

Acknowledgments. I would like to thank Dr. Gürol Baba for his helpful comments and suggestions to improve this paper.

Received: 04/03/2018. **Accepted:** 01/10/2018

Searle, the Turing Test and Strong AI

After almost four decades of its publication, John Searle's Chinese room thought experiment (1980) still puzzles the field of artificial intelligence (AI). Using this thought experiment, Searle convincingly – but mistakenly in my opinion – defended that no matter how complex and well programmed a computer performing symbol manipulation is at present, and could be in the future, it cannot think in the manner human beings do. He also rejected Alan Turing's (1964) claim that the Turing test is a sufficient condition for determining whether an AI system really thinks. Although the specifics of how the Turing test should be performed remains a controversial topic (Traiger 2000), the test can roughly be described as follows:

> There is a computer and two humans. One human is the interrogator. She or he communicates with the computer and the other human using a teletype or computer terminal. The interrogator is told that one of the two individuals it is communicating with is a computer, and that the other is a human. The computer's goal is to fool the interrogator into thinking that it is the human. The other human's goal is to convince the interrogator that he or she is the human. The computer attempts to achieve the deception by imitating human verbal behavior. If an interrogator does not make the right identification, where a "right identification" is identifying the computer as a computer, then the computer passes the test. (Traiger 2000: 561)

According to Turing, if a computer can convince an interrogator that it is a human being as frequently as a human being can, the computer should be considered to be a thinking being or to possess human-like cognitive capacity. He further asserts that the other philosophically complicated concepts of *thinking* are vague and useless because they are not testable or verifiable, and that there is no acceptable concept of thinking to replace his behavioristic concept of thinking based on Turing test results (Turing 1964).

After a decade, Roger C. Schank and Robert P. Abelson (1975) were working on a computer program (script applier mechanism [SAM]) capable of inferring implicit propositions in natural-language stories. The implicit propositions, which could be inferred by a human quite easily, were not logically necessary conclusions of the explicit statements in the stories. The following story is an example that SAM analyzed and answered the questions about: "John went to a restaurant. The hostess seated John. The waiter came to the table. John ordered lobster. John was served quickly. John left a large tip. John left the restaurant" (Schank and Abelson 1975: 153). Although the reason why John left a large tip is not explicitly stated, SAM was able to deduce that the probable reason was the quick service (Schank and Abelson 1975: 154). At the time, Schank and Abelson (1975: 155) estimated that with some improvement, their program (SAM) could *understand* simple stories about a range of domains.

In response to Turing (1964), Schank, and Ableson (1975); Searle (1980) posited that in the absence of a foundational scientific and/or engineering revolution that enables computers to perform tasks beyond symbol manipulation, no computer including ones that would pass Turing test can perform human-like thinking. He claims that insignificant developments, like coding a software aiming to pass behavioral tests like the one Turing proposed, overlook one of the core concepts of the philosophy of mind: *intentionality*.

In the Chinese room thought experiment, Searle imagines himself acting as a computer that is trying to understand Chinese stories, where Searle has no knowledge of the Chinese language. He is placed in a room with some syntactic instructions (algorithm) in English that will help him to manipulate Chinese symbols properly. Next, some Chinese stories and Chinese questions about these stories are passed to Searle from outside the room and he tries to answer the questions in Chinese with the help of English instructions. The algorithm is so comprehensive and Searle is so skillful in applying it that he manages to prepare correct Chinese-language answers to the questions quickly, although he does not know the language. If all these are true, a Chinese-speaking individual who does not know what is happening in the Chinese room could possibly think that the person in the room understands Chinese. That is to say, Searle would be able to pass the Turing test in Chinese, without understanding anything about the stories, questions and his answers to the questions. In brief, this thought experiment shows that a person or a computer with no understanding, can manipulate symbols (Chinese letters) meaningfully, with the help of an algorithm. Accordingly, it is possible to pass the Turing test without understanding or thinking like a human being.

This is a sound argument: The Turing test does not account for phenomenological, or at least the intentional aspect of thinking. Nevertheless, Searle bases a stronger and controversial conclusion on the Chinese room: no machine based on computational symbol manipulation can perform human-like thinking. Needless to say, this is a negative existential statement, proof of which is more demanding than the one above. After all, in the Chinese room, Searle tests only a particular kind of algorithm for a particular kind of problem.[1] Searle would not make the mistake of relying on the following invalid argument in reaching his stronger and much more controversial conclusion:[2]

1) In the room Searle acts as a computer that manipulates symbols in order to communicate in Chinese.
2) Searle does not understand or think about content of Chinese symbols.
3) No computer using pure symbol manipulation can understand or think (which does not follow from 1 and 2).

Apparently, Searle uses at least one additional premise to show that (3) is true. One such premise that leads Searle to conclude (3) is that the person in the Chinese room has access to all the necessary tools that can be used by a computer for the task of understanding. Another of Searle's additional premises is that there exists no better algorithm for

[1] See Jerry Fodor's comment in Searle 1980.

[2] Searle seems to base his thesis on a clearly invalid argument in the following passage: "[…] I offer an argument that is very simple: instantiating a program could not be constitutive of intentionality, because it would be possible for an agent to instantiate the program and still not have the right kind of intentionality (Searle 1980: 450-451)." I certainly do not see how the possibility of an agent's instantiating a program that does not have the right kind of intentionality show the impossibility of an agent's instantiating a program that does have the right kind of intentionality.

understanding Chinese than the one the person in the Chinese room is given. And finally, no matter how long Searle stayed in the room, he could not start to understand Chinese. Without these additional premises, the Chinese room thought experiment cannot derive his controversial conclusion that no computer can perform human-like thinking. Unless he justifies these additional premises, a natural objection to Searle could be that even if the person in the room did not understand Chinese, if he used a better algorithm to communicate and understand Chinese, or if he had access to other tools a computer could, or if he had some more time experiencing symbol manipulation, he might have very well been able to understand Chinese. To prevent this objection, his argument suggesting that a computer cannot perform human-like thinking should be in the following form:

1) In the room Searle acted as a computer that manipulates symbols in order to communicate in Chinese.
2) (No matter how long Searle stayed in the room) Searle *cannot* understand or think about content of Chinese symbols.
3) No computer could have a better algorithm or access to more useful tools for understanding or thinking about content of Chinese symbols than Searle in the Chinese room.
4) No computer can understand or think (from 1, 2 and 3).

Intuitively, Searle seems correct in that the person in the Chinese room would not have accurate intentional states related to Chinese symbols at least in a short while. That is to say, the person would be unaware of how the symbols are connected to the world outside the room. On the contrary, this paper argues that (2) and (3) are false. That is to say, given enough time, Searle in the room could actually understand Chinese stories to some extent. Moreover, if he were given access to some additional tools which a computer could have access to, he would understand Chinese stories much easier, faster and more in depth.

As Searle points out, *at the beginning*, the person in the room would lack intentionality in the sense of directedness. Directedness, in its broadest definition, is the ability to establish relationships between mental and external objects. Both a standard computer and I may express the sentence, “The Moon is Earth’s natural satellite;” however, unlike me, a standard computer does not associate this statement with the two celestial objects. As far as a standard computer is concerned, the word Moon does not refer to a celestial body; therefore, it would be erroneous to assume that the computer understands that the Moon is Earth’s satellite. This problem is related to the difference between sentence and proposition or content and symbols. The central question concerning understanding is whether it is possible for semantics to emerge only from syntactical manipulation. Searle thinks it is not:

> Computation is defined purely formally or syntactically, whereas minds have actual mental or semantic contents, and we cannot get from syntactical to the semantic just by having the syntactical operations and nothing else. To put this point slightly more technically, the notion “same implemented program” defines an equivalence class that is specified independently of any specific physical realization. But such a specification necessarily leaves out the biologically specific powers of the brain to cause cognitive processes (Searle 2010: 17).

Could Searle be correct in his claim that there are some biological/physical causal properties of the human brain that render it something more than a computational symbol manipulator? In this paper, the puzzling concepts of "meaning," "reference," "intentionality" and relations between them are not investigated thoroughly, therefore this question is not answered conclusively. This paper mainly aims to show that Searle's Chinese room did not succeed to show that a computer cannot think, functionalism is false and semantics cannot emerge on syntactical operations. To do that, first I will modify the Chinese room thought experiment in a way that the person in the room is given more data and a proper algorithm for decoding Chinese, and consequently, these modifications will enable the person in the room to start understanding Chinese.

An Unfair Version of the Chinese Room (UVCR)

Proponents of robot reply underline the fact that in the Chinese room thought experiment, the person in the room does not have access to the external sensory data that a computer may have via a camera and/or a microphone and defends that this is the main reason why Searle in the room does not understand Chinese (Searle 1980: 431). The unfair version of the thought experiment (UVCR) that I will describe in this section, will intuitively support that robot reply is correct in that some set of relations between sensory data and syntax could help to bridge the gap between syntax and semantics. However, as its name suggests, this version of the Chinese room cannot be conclusive, as it gives the person in the room an unfair advantage that a computer does not possess.

In UVCR a person enters the room with English-language instructions (algorithm), as in Searle's version, but now the algorithm, the Chinese stories and questions he receives are accompanied by visual data concerning the meaning of the Chinese characters. For example, for the Chinese sentence, "苹果 是 红色 的," which means "the apple is red," some additional data are provided, such that:

1. Pictures of apple in different color accompany the syntax "苹果" (apple);
2. The expression 红色 (red) is accompanied by examples of red objects;
3. Finally, the sentence "苹果 是 红色 的" is accompanied by a picture of a red apple and graphical data indicating the subject of the sentence is the apple and predicate is to be red.

Provided these additional data, Searle (a person) in the room would not only pass the Turing test, but would also begin to understand Chinese. This time, the person in the room would have accurate intentional states about the given Chinese symbols. In other words, when the person in the room reads the sentence "苹果 是 红色 的" they would associate it with an object (the apple) and a predicate (to be red).

It is now clear that if the person in the room were provided some additional tools or data that a computer could have, they could understand Chinese. At first glance, providing some visual data about the world to the person in the room may not seem unfair (despite the fact this paper argues otherwise), since, via a camera, a computer can also collect data about the world. If UVCR were to be fair, it would show that there is no sound argument *put forward by Searle* to believe that a properly programmed and adequately configured computer identifying relationships between syntax and the world cannot think. On the other hand, UVCR provides an unfair advantage to the person in the room: After all for the person in the room, the pictures that are associated with Chinese words are not mere symbols to be manipulated; they are symbols with content. That is to say, before entering the room, Searle already possessed the idea of what an apple is and how it seems. He simply associated the content of “apple,” which he had already possessed, with the syntax of “苹果,” as opposed to building the content of “apple” via mere symbol manipulation. Searle’s main conclusion that semantics cannot emerge through symbol manipulation cannot be debunked by UVCR as it is. However, this paper argues that the original version of the thought experiment is also unfair to the person in the room, as it did not allow Searle to perceive the outside world, which is crucial for the efficiency and quality of understanding the process as UVCR suggests. The bottom line is that Searle’s preference to hypothesize a person with a semantic history in the room to test strong AI creates a dilemma. If we were to allow the person in the room to access perceptual data about the Chinese symbols, we fail to test whether semantics can emerge through mere symbols; and, if we do not, we are unfair to the person in the room (or computers), since perceptual data is an important (although, not necessary, as it will be defended later) component of understanding in an ordinary sense.

The common sense concept of understanding requires an accurate establishment of relationships between syntax and the world, i.e., the meanings of words. Individuals learn the meaning of a word by interacting with the external world and constructing relationships between words and their correspondences in the world. A person or computer manipulating syntax in the absence of sense data cannot establish these substantial relationships, therefore lacks intentionality and understanding in an ordinary sense. This is one of the reasons why I conclude that Searle’s original Chinese room thought experiment is set up in a way that is unfair to the person in the room, computers, and strong AI. It does not allow the person in the room to perceive the world. On the other hand, modifying his thought experiment as it is done in this section, only makes it unfair in the opposite sense, as UVCR lets the person in the room use their semantic background to make sense of Chinese symbols. The next section constructs a conclusive version of the thought experiment, which decisively proves that Searle’s argument against strong AI is not valid.

As a final remark, notwithstanding sense data’s help for fast, intuitive and in depth understanding, this paper defends that availability of sense data is not a necessary condition for understanding, which will also be defended in the following section.

A Fair Version of the Chinese Room (FVCR)

Before presenting the fair version of the Chinese room, human capacity of pattern recognition will be elaborated briefly. Humans cope with an incredibly complex world. According to Claude Edward Shannon’s (1988) conservative calculations, in a chess game, there are 10^{120} variations (each variation is a complete possible chess game) to deduce the best possible starting move. “A machine operating at the rate of one variation per micro-second would require over 1090 years to calculate the first move! (Shannon 1988: 4)” It is practically impossible even for computers to process these data deductively. On the other hand, we humans do not have to deduce all variations to make a “good” decision, thanks to our pattern recognition ability. Human players categorizes chess entities like open file, fork, pin, mid-game, end-game, Slav formation, etc. and learn or discover advantageous chess moves in this conceptual space, which is much less complex than the well-defined variation space. In short, instead of processing all variations, humans categorize them and act according to which category they fall under. Recognizing patterns and making inferences about categories is more economical, and mostly only possible option with respect to cognitive resources. Now the question is, how do computers play chess if the game is incredibly complex? They also use pattern recognition techniques and they can be comparable to (and even better than) humans recognizing patterns at least in the domain they are designed for (Rasekhschaffe and Jones 2019). Cristopher M. Bishop (2006: 1) describes the aim of computational pattern recognition as the “discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories,” and underlines the fact that humans also have that capacity.

Even if chess is one of paradigmatic examples of complex systems, need and use of pattern recognition ability is not limited to challenging practices like playing chess or doing science. Apparently simple tasks, like perceiving the environment, walking, speaking, reading, driving, and so on, take place in complex systems, and human beings use their ability to recognize patterns in these domains too. To illustrate, while driving, an expert driver tend to recognize patterns of the engine sound and act accordingly, without following well-defined rules like “shift to 2nd gear when the speedometer needle points to 10 (Dreyfus 2004: 177).”[3]

Why Searle holds the view that the person in the Chinese room or a properly programmed computer would not be able to understand is his overlooking this essential capacity of pattern recognition and this paper will clarify how it is possible for the person in the Chinese room to understand Chinese, given that s/he has a moderate capacity to recognize patterns in the following thought experiment.

In the final, fair version of the Chinese room (FVCR) the person entering the room is an alien, unfamiliar with Earth or any other planet. She has always lived in a starship with no windows that would have allowed her to see the space. The starship is governed by artificial laws of physics which are different than ours. The alien has never seen an apple, a car, the Sun, the Moon, and so on. The alien lives in an environment which possesses entirely different characteristics than Earth. Naturally, she is wholly unfamiliar with the English language. Then she enters the “*English* room” in her starship. Needless to say,

[3] For a comprehensive discussion of human ability to cope with complex environments, see Dreyfus 1972; Dreyfus and Dreyfus 2000; Dreyfus 2004. On the other hand, note that I do not agree Dreyfuses in that non-representational learning cannot be simulated by computers.

she is provided with certain instructions in her native language that will enable her to manipulate English letters correctly. In FVCR the alien is provided with English stories, questions and simple pictures associated with the given English words from the stories and questions, and she manipulates English symbols to write down reasonable answers to the questions. Let us assume that these stories focus extensively on simple information about the Solar System. For the alien, these pictures and letters are nothing more than meaningless symbols, as she is wholly ignorant of the referent of pictures and English words. With the help of instructions she is given, the alien is able to manipulate the English symbols accurately without any understanding about the content of the sentences she reads or writes. However, in this thought experiment, the alien enjoys manipulating symbols to the extent that she remains in the room for months.

The diligence of the alien brings about a twist in the thought experiment: After seeing thousands of texts about the Solar System, she begins to realize that there are certain patterns in the strings of symbols in the English stories. She discovers, for example, that while the string of “Venus,” “Earth” or “Saturn” concludes with “is a planet;” the string of “the Moon,” depicted in a similar way to these planets, concludes with “is not a planet.” or “is a satellite of the Earth” After investigating hundreds of sentences related to planets and natural satellites, she creates two concepts one corresponding with our concept of “planet,” and a second corresponding to our concept of “satellite”. It cannot be easily be claimed that it would be possible for her to construct a concept of planet or satellite that is similar to ours. After all, her concepts, let us call them “planet′” and “satellite′” are constructed in the absence of detailed information concerning planets or natural satellites. She knows, on the other hand, that planets′ or satellites′ are in some way related to circular shapes in different sizes and surfaces. Another thing she notices is that circular shape a satellite′ is associated with is always smaller than that of the planet′ it orbits′ and mostly those of other planets′. She further discovers that every string that accurately concludes with “is a satellite.” also concludes with “is satellite of [a planet].” In other words, every satellite′ has a certain relationship with a planet′. In the same way, she *understands* that there is a relation, namely “orbiting around” between planets′ and the Sun′, and satellites′ and planets′. Moreover, she notices that every satellite′ is a satellite′ of the planet that it is “orbiting around′.” She suspects that being satellite of′ is equivalent to orbiting around′. Finally, she discovers that whatever the relation of “being bigger than” is, it is an *order relation*. That is to say, for A, B and C are different from one another, if A is bigger than B and B is bigger than C, then A is bigger than C. Consequently, upon reading that the Sun is bigger than Mars and that Mars is bigger than a meteor, she can now conclude that Sun is bigger′ than a meteor. She does not stop there, and sees that if A is bigger than′ B, the circular shapes provided for the strings representing A is bigger than those of B. Now, she accurately believes that she understands the meaning of “being bigger than”. She understands the meaning of “bigger than” as a human being would do in their childhood and she has the appropriate intentional states concerning this concept. She understands the meaning of “being bigger than” and this understanding does not stem from past experiences collected outside the English room. Note that some of the understandings provided above does not require simple pictures associated with English words. The pictures which are analogous to sense data are not a necessary condition for understanding but a helpful tool for faster, more comprehensive and deeper understanding.

Now in FVCR has the alien begun to understand English? Apparently, she has a degree, scope and depth of understanding of English texts like any of us. For now, the alien's understanding is limited to a very small domain but whatever she understands, she understands in the same way we do.

Note that FVCR reveals only one possible outcome about the English room. Therefore, FVCR does not show that any alien would start understanding in the English room. It is true that depending on the instructions the alien is given and pattern recognition capabilities and cognitive tendencies of her, she could get confused in the room and may not understand anything about English symbols, no matter how long she stays. To illustrate, she may wrongly assume that pictures associated with English words are not representations of sense data signified by the words but they themselves are some additional signifiers, or her memory or pattern recognition capability may not be enough to see the relationships in complex texts. On the other hand, FVCR shows that some aliens which have necessary pattern recognition capabilities and motivation would start understanding in the room. Demonstration of this possibility is enough to negate Searle's view that strong AI is false.

It may be claimed, on the other hand, that what the alien learnt and understood in the English room is all about the Latin symbols and syntax of English language, not about the world. Assuming that she does not have access to pictures associated with English words, this claim would be even more appealing. After all, for the alien, the string of "planet" does not signify the celestial objects we call "planets". Yet, this is not because she does not construct concepts about the entities in the world, which strings of Latin symbols could possibly refer to but because her concept of planet′ is much less complete than that of ours. Throughout the time the alien spent in the room, she has been constructing a concept of planet′ reference of which she has very little knowledge, just as we do when we clumsily construct a concept of a "wave function" while reading an advanced article in quantum physics with almost no prerequisite knowledge. In short, I assume that the alien in the English room knows that these Latin symbols are meant to express propositions about the world, just as the Searle in the Chinese room does know that Chinese symbols are related to the world. Someday, let us say she arrives on Earth and begins to perceive our world including the planets, the satellites and the Sun within our solar-system. In this case, she would start to deepen her understanding of the English language and our world far more progressively and convergent to our understanding. This article defends that FVCR shows that Searle is mistaken in that he would not understand Chinese in the room. Moreover, I suspect that FVCR may not be necessary to show that Searle is mistaken. It can be argued that with or without English instructions, the person in the original Chinese room could, in principle, understand Chinese, if they are given enough time and, hence, the necessary experience about meaningful Chinese texts. This conclusion is evident in the fact that while new-born babies perceive the world via meaningless symbols, they somehow make sense of the world without instructions on how to manipulate them. They naturally experience reality, recognize discernible patterns in it and map the meaningless symbols onto the world. If a new-born baby can understand a new language without any instructions, then it should be possible for Searle in the Chinese room to understand Chinese with or without instructions.

Another question in need of an answer is how is the task assigned to the person in the Chinese room related to the strong AI, which is the thesis that a properly configured and programmed computer can think in the same way we do? This article defends that a person's success in understanding Chinese in the Chinese room is neither a necessary, nor a sufficient condition for strong AI: It is not a sufficient condition: In this section it is shown that a person with human cognitive capacities could pass the English-room-test (or the Chinese-room-test); yet, this does not conclusively show that a computer could accomplish the same task by using pure symbol-manipulation, since we do not agree on the premise that all cognitive capacities of humans are based on computational symbol manipulation. To clarify, an alien could learn a language solely by manipulating symbols; but while she is doing so, she is using various cognitive processes (consciousness, various reasoning methods, experiencing qualia, pattern recognition, and so on) and it is not obvious that these cognitive skills could all be replicated by algorithms on symbol manipulation. Assuming that all of a person's cognitive capacities are based on symbol manipulation to show that a computer working on symbol manipulation can think in the same way a human does, would suffer from a circularity problem.

On the other hand, a person's success in understanding Chinese in the room is certainly not a necessary condition for strong AI either: There are numerous ways (algorithms) of manipulating symbols, and the person in the Chinese room uses only one particular way of it, which is defined by "the instructions in English." Even if we agreed that it is impossible for an ordinary person in the Chinese room to understand Chinese with one specific algorithm and one set of cognitive skills s/he possess, this does not provide conclusive evidence that no person in the room can understand Chinese regardless of the instructions s/he follows or the cognitive skills s/he possesses. Accordingly, Searle's version of the Chinese room, and his premise that the person in the room could no way understand Chinese, does not show that no computer can think no matter what algorithm it uses and how it is configured. After all, a person in a "Fibonacci room" with no knowledge of Fibonacci numbers cannot calculate Fibonacci numbers, if s/he is given an inaccurate set of instructions but this does not show that no computer can calculate Fibonacci numbers, no matter which algorithm it uses. Just like it is still possible that a properly programmed computer can calculate Fibonacci numbers, it can still be possible that properly programmed computer can understand just as humans do.[4]

Therefore, the Chinese room thought experiment is not directly related to strong AI thesis. The Chinese shows only that a computer's passing the Turing test does not guarantee that it thinks and understands in the way humans do.

[4] See Churchlands' (1990: 35) luminous room argument in response to the Chinese room.

Searle's Response to the Robot Reply

I defend that even without perceptual data, a person with pattern recognition capability, could, in principle, begin to understand a foreign language that they are manipulating. This is where the position of this paper differs from "robot reply," or any understanding which holds the view that perceptual (sense) data is necessary for understanding. However, the availability of perceptual data would enormously boost degree, scope and depth of their understanding. This is why, in this section, I will discuss "Robot Reply" in relation with human/computer capacity of pattern recognition.

Like Jerry Fodor[5], I find Searle's response to *robot reply* unconvincing (Searle 1980: 431):

> [T]he addition of such "perceptual" and "motor" capacities adds nothing by way of understanding, in particular, or intentionality, in general, to Schank's original program. To see this, notice that the same thought experiment applies to the robot case. Suppose that instead of the computer inside the robot, you put me inside the room and, as in the original Chinese case, you give me more Chinese symbols with more instructions in English for matching Chinese symbols to Chinese symbols and feeding back Chinese symbols to the outside. Suppose, unknown to me, some of the Chinese symbols that come to me come from a television camera attached to the robot and other Chinese symbols that I am giving out serve to make the motors inside the robot move the robot's legs or arms. It is important to emphasize that all I am doing is manipulating formal symbols: I know none of these other facts. I am receiving "information" from the robot's "perceptual" apparatus, and I am giving out "instructions" to its motor apparatus without knowing either of these facts. I am the robot's homunculus, but unlike the traditional homunculus, I don't know what's going on. I don't understand anything except the rules for symbol manipulation. Now in this case I want to say that the robot has no intentional states at all; it is simply moving about as a result of its electrical wiring and its program. And furthermore, by instantiating the program I have no intentional states of the relevant type. All I do is follow formal instructions about manipulating formal symbols (Searle 1980: 420).

First, I agree with Searle in that the robot version is not fundamentally different from the original thought experiment; but, unlike Searle, I defend that in both versions, given enough time, a human being would begin to make sense of the symbols. More precisely, with the lines of thinking in the FVCR, Searle in the Chinese room would begin to. It is evident, with the line of thinking in the FVRC, Searle in the Chinese room connected to a robot or Searle in my skull connected to my nerves could begin to recognize patterns in the symbols he is manipulating, discover relationships in them and consequently understand the meaning of symbols (or signals.) Even though Searle, when in the room connected to the robot does not perceive the outside world directly, he is fed by Chinese characters representing the sensory data coming from the camera attached to the robot.

[5] I do not agree with Fodor's initial comment on the Chinese room stating that "[I]nstantiating the same program that the brain does is not, in and of itself, a sufficient condition for having those propositional attitudes characteristic of the organism that has the brain. If some people in Al think that it is, they're wrong (Searle 1980: 431)." See also Fodor 1991.

After processing these symbols, Searle sends back Chinese characters that manipulate the motors which move the robot. As far as we know, this is remarkably similar to how a human brain works. It perceives the world by electrical signals (symbols) and acts on outside signals by again using electrical signals. When the Searle in the Chinese room receives the Chinese characters representing the sensory data of an apple (without knowing that the syntax represents the apple), he might question whether this representation has a certain relationship (being in color) with syntax representing sense data of "red" or "green". After all, Searle has previously manipulated and recognized patterns in countless texts containing Chinese symbols representing sense data of red and green apples that the robot he is connected interacted with. So, Searle by sending certain symbols to the robots visual components, may check if it is red or green (more precisely, if it has certain relationship with symbol patterns representing red or green)If the apple is green, Searle (ignorant about the real nature of apple, green or red), might decide to send certain symbols to obtain a (metaphorically speaking) pleasurable string of symbols, and this way, would cause the robot to eat the green apple and direct the robot to enter into a *goal state*.

It is true that Searle in the Chinese room connected to a robot would also be unaware what he is doing in an ordinary sense at least, at the beginning. He would be unaware that the apple is something to be eaten and its color is an electromagnetic property of the apple. He would be unaware of many things that we know about apples, because he is not receiving the same symbols that we do, and he is not manipulating the symbols via a similar mechanism that a human brain does. However, he would begin to understand symbols he is manipulating. and by time, in the same way as the alien in the English room, his understanding would be sharpened And after a while, perhaps, he would be aware of some facts that we are not aware about apples, the color of green and red (again because he is not fed by the same data we are fed with and algorithm he follows is different from the one our brains does). Moreover, I admittingly speculate that if the Searle's mechanism of manipulating the symbols were similar to that of a brain, how he understands the world would eventually converge with our way of understanding.

Conclusion

In essence, Searle's argument against functionalism, the robot reply and strong AI ignores a capability that both human beings and properly programmed computers share: pattern recognition. Human beings are capable of capturing patterns in complex inputs, consciously and unconsciously. In recognizing patterns, our nervous systems (mostly unconsciously) filter insignificant variables and allow us to make sense of complex electronic impulses that represent the world. Accordingly, pattern recognition is one of the tools we use to invent or discover meaningful higher-order concepts hidden in meaningless symbols (like Chinese letters or electrical signals). Meaning and understanding concerns these patterns hidden in these incredibly complex electrical signals (or Chinese letters in the Chinese room) coming from our sense organs.

The Chinese room, as it is, does not disprove strong AI, as shown in the fair version of the Chinese room: it is possible for the person in the room to understand Chinese if they are provided with enough time and possess a moderate capacity to recognize patterns emerging in the symbols they manipulate. Understanding Chinese in the room would be even easier and faster if the person in the room were fed by (familiar or alien) audiovisual data coming from outside, since our brains have specifically evolved to recognize patterns in these kinds of sensory data. Arguably, two of the "mysterious" causal powers of the brain that puzzles Searle are 1) the brain's pattern recognition capacity and 2) the brain's capacity to construct relationships between sets of symbols (as our brains do when we relate the word "table" to some visual data belonging to a table).

Intentionality, artificial pattern recognition and artificial concept creation are central issues for strong AI. Mechanism(s) by which meaningless sensory data result(s) in human-like understanding/thinking continue to remain a mystery; however, Searle fails to provide any convincing evidence that the brain is the only physical structure capable of human-like thinking or that the causal relationships formed by the brain are the only possible relationships that could provide the foundation for thinking to emerge. I believe that intentionality can be reduced to a set of well-defined functions, which can be realized in various types of hardware composed of different materials, including the brain, computer hardware or any other structure providing the opportunity to represent and manipulate dynamic complex relationships. Provided that algorithms that efficiently recognize patterns, create concept, and bind the concepts to the world could be constructed, I do not see why computers categorically may not think.

References

Bishop, C. M., 2006. *Pattern Recognition and Machine Learning.* New York: Springer-Verlag.

Churchland, P. M., Churchland, P. S., 1990. Could a.Machine Think? *Scientific American* 262(1): 32–37. https://doi.org/10.1038/scientificamerican0190-32

Dreyfus, H. L., 1972. *What Computers Can't Do.* New York: Harper & Row.

Dreyfus, H. L., Dreyfus, S. E., 2000. *Mind over machine.* New York: The Free Press.

Dreyfus, S. E., 2004. The Five-Stage Model of Adult Skill Acquisition. *Bulletin of Science Technology & Society* 24(3): 177–181. https://doi.org/10.1177/0270467604264992

Fodor, J. A., 1991. Afterthoughts: Yin and Yang in the Chinese Room. In D. M. Rosenthal (ed.), *The Nature of Mind*, New York: Oxford University Press, pp. 524–525.

Rasekhschaffe, K. C., Jones, R. C., 2019. Machine Learning for Stock Selection. *Financial Analysts Journal* 75(3): 70–88. https://doi.org/10.1080/0015198x.2019.1596678

Schank, R., Abelson, R., 1975. Scripts, plans and knowledge. *The Proceedings of Fourth International Joint Conference on Artificial Intelligence*, Tblisi, pp. 151–157.

Searle, J. R., 1980. Minds, Brains, and Programs. *The Behavioral and Brain Sciences* 3(3): 417–457.

Searle, J. R., 2010. Why Dualism (and Materialism) Fail to Account for Consciousness. In R. E. Lee (ed.), *Questioning Nineteenth Century Assumptions about Knowledge, III: Dualism*, New York: SUNY Press.

Shannon, C. E., 1988. Programming a Computer for Playing Chess. In D. N. Levy (ed.), *Computer Chess Compendium*, New York: Springer-Verlag, pp. 2–13. https://doi.org/10.1007/978-1-4757-1968-0_1

Traiger, S., 2000. Making the right identification in the Turing test. *Minds and Machines* 10(4): 561–572.

Turing, A. M., 1964. Computing machinery and intelligence. In A. R. Anderson (ed.), *Minds and machines*, New Jersey: Prentice Hall, pp. 4–30.