25 aniversario

ARTÍCULOS

UTOPÍA Y PRAXIS LATINOAMERICANA. AÑO: 25, nº EXTRA 5, 2020, pp. 179-186
REVISTA INTERNACIONAL DE FILOSOFÍA Y TEORÍA SOCIAL
CESA-FCES-UNIVERSIDAD DEL ZULIA. MARACAIBO-VENEZUELA
ISSN 1316-5216 / ISSN-e: 2477-9555

The Chinese Room Argument in the Context of the Rational Action Theory

Argumento de la sala china en el contexto de la teoría de la acción racional

Maxim Aleksandrovich MONIN
https://orcid.org/0000-0001-7304-0896
monin.maxim@gmail.com
Sechenov First Moscow State Medical University (Sechenov University), Russia

Vera Albertovna TEREKHOVA
https://orcid.org/0000-0002-7029-7888
ter.vera@mail.ru
Sechenov First Moscow State Medical University (Sechenov University), Russia

Elena Vladimirovna LEDENEVA
https://orcid.org/0000-0001-9867-369X
ledenevaelena72@mail.ru
Sechenov First Moscow State Medical University (Sechenov University), Russia

Este trabajo está depositado en Zenodo:
DOI: http://doi.org/10.5281/zenodo.3984228

RESUMEN

El artículo describe un conocido experimento mental de J. Searle llamado la Sala China dirigida contra las teorías de la inteligencia artificial fuerte en el contexto de una larga polémica tras la publicación del artículo de Searle, así como su libro posterior "Racionalidad en acción" (2001). Su separación radical en el marco de su experimento mental sugerido fue la más criticada ya que se observó una simplificación que no cumplía con las teorías reales de la inteligencia artificial. El libro de Searle publicado en 2001 deja en claro que el filósofo significa ante todo la diferencia dentro de la semántica misma: conectada principalmente con la acción y la sintaxis.

Palabras clave: Prueba de Turing, sala china, intencionalidad, sintaxis, semántica, subcognición**.**

ABSTRACT

The article describes a well-known thought experiment of J. Searle called the Chinese Room directed against theories of strong artificial intelligence in the context of longstanding polemic following the publication of Searle's paper as well as his later book "Rationality in Action" (2001). Their radical separation in the framework of his suggested thought experiment was criticized the most as there was seen simplification which did not comply with real theories of artificial intelligence. Searle's book published in 2001 makes it clear that the philosopher means first of all the difference within the semantics itself: connected mainly with action and syntax.

Keywords: Turing test, chinese room, intentionality, syntax and semantics, subcognition.

Recibido: 24-06-2020 ● Aceptado: 20-07-2020

INTRODUCTION

Turing test and Searle's experiment

About forty years ago (in 1980) an American philosopher John Searle published in his paper "Minds, Brains, And Programs" (Searle: 1980) his famous refutation of what he called strong AI (artificial intelligence) thesis, which as stated by Searle claims "that the appropriately programmed computer literally has cognitive states and that the programs thereby explain human cognition" (Searle: 1980, 417). The direct addressee of Searle's paper, as he writes, is the research of R. Shank and R. Abelson (Shank, Abelson: 1977, 248) whose authors claim that they managed to create a computer program capable of understanding the meaning of human stories. For example, regarding a story: "a person went to a restaurant and ordered a hamburger; when the hamburger was brought and it turned out to be burnt, the man left the restaurant in indignation without having paid for the order". The question was: "did he eat the hamburger?" The "appropriately" programmed computer responded that most likely not. In his article Searle does not analyze either the procedure for testing computers used by Shank and Abelson or the principle of their program's operation. He raises a question whether it is possible at all to speak of comprehension when a computer without possessing corresponding visual, olfactory, taste experience cannot know the meaning of words "hamburger", "burnt", etc. As fairly believed by Searle, such AI research as that carried out by Shank and Abelson follows the paradigm of A. Turing's well-known test according to which a satisfactory imitation of "human answers" by a computer is similar to reasonable answers of a person. In the Turing test a person playing a role of an expert asks questions in a hard copy format and in the same way receives answers from two interlocutors invisible to him, one of whom is a person, the other one is a specially programmed computer. According to Turing, the criterion for satisfactory passing the test was identifying the computer by the expert in no more than 70% of cases after a five-minute survey (Turing: 1950, 441), which in Turing's opinion would allow to believe that the computer is capable of thinking.

In his refutation of strong AI thesis Searle suggests his own interpretation of the Turing test, which Searle in common with Turing considers actually feasible for a computer. That said, Searle discards those reservations which Turing makes regarding the procedure of the test, being guided by the thought that even if passing the test by a computer is perfect from the point of view of an external observer it is not just far from being similar to human thinking but has nothing in common with it. The thought experiment proposed by Searle, his famous Chinese room, which some researchers call the "parody" (French: 2000, 660) of the original Turing test is rather its strictly inverted form. In Searle's variant the computer is also interrogated, however an expert here is not the questioner but a person who is "on the other side" of the wall and performs functions of a "processor" of the computer: the person receives some strange symbols from the outside (the analogue of an incomprehensible to the computer ordinary language in the experiment is an incomprehensible to a "normal" English-speaking individual Chinese language), with the help of some program he relates them to other symbols, the meaning of which is also unknown to him, and gives the questioner a reply meaningful from the point of view of the latter. However, from the point of view of the expert - "the processor" - he does not perform any mental action. Moreover, Searle continues, no matter for how long he will be manipulating the Chinese hieroglyphs, he will never be able to speak Chinese language or understand the meaning of his activity.

Searle's thought experiment is supposed to illustrate the key idea for his refutation of strong AI thesis that "programs are formal (syntactic)" whereas "human minds have mental contents (semantics)" (Searle: 1990, 27), and that "syntax by itself is neither constitutive of nor sufficient for semantics". Accordingly, a logical conclusion of the three given premises is a statement that programs do not create consciousness and are not sufficient for it. This negative conclusion Searle complements with two positive ones: (1) consciousness, unlike a computer program, is always directed towards an object, i.e. it has intentionality (Searle: 1980, 436) and (2) consciousness always has neurobiological rootedness: «as far as we know, every "mental" event, ranging from feelings of thirst to thoughts of mathematical theorems and memories of childhood, is caused by specific neurons firing in specific neural architectures» (Searle: 1990, 29), while computers “are completely independent of any reference to a specific type of hardware implementation”.

RESULTS AND DISCUSSION

Considered False but not Disapproved of

Despite the fact that Turing’s thesis of strong AI caused considerable criticism at one time (a significant part of Turing’s article in 1950 contains the latter’s answers to the raised by that time objections about possessing by a computer such features as imagination, creativity, emotions, ability to learn, etc.) and also despite the fact that no convincing confirmation has ever been received for the Turing hypothesis that after 50 years (i.e. by the year 2000) the machines will be able to pass the Turing test, Turing's ideas on AI are evaluated in modern scientific literature quite benevolently (Saygin, Cicekli: 2000, 464-465). The same cannot be said about the experiment with the Chinese room suggested by Searle. An article by Searle in 1980 was published along with twenty-six critical responses to it, to the majority of which Searle responded with a new article (in 1990). Accordingly, in subsequent publications on the Chinese room these “primary comments” and Searle’s responses to them were added to the original text by Searle. Critics tried to refute the idea contained in Searle's thought experiment, referring to the fact that even if the expert in the Chinese room does not understand Chinese but the system as a whole (i.e. the whole room: expert + program + information entry and exit procedures, etc.) does (The Systems Reply). Searle replied that even if the expert memorizes the whole program and becomes the room himself, there will still be no comprehension. Another objection concerning the fact that if the expert is placed inside a robot with video cameras and sensors, the expert will perceive what is happening around and his manipulations will become conscious (The Robot Reply). Searle replied that even in this case the expert will not be able to relate the symbols transmitted to him to his perception experience, which is to say he will still not be able to comprehend them. Similarly, Searle responds to the objection that the program can simulate the brain of a Chinese speaking person (The Brain-Simulator Reply) – theoretically such brain simulation can consist of anything, for example, water pipes and rolls of toilet paper, and its manipulation with Chinese symbols will still not be able to go to the level of semantics comprehensible for the expert. Finally, to the objection that such program will work too long, Searle suggests to put in the Chinese room any amount - even a cool billions of people who do not speak Chinese; the result, he said, would be the same: none of them will ever comprehend the meaning of their work or start speaking Chinese.

One can see that the abovementioned and similar objections to Searle could be more or less convincingly neutralized introducing corresponding changes into the original concept of the Chinese room. But at the same time, the very idea of the Chinese room, which from the very beginning had a status of an exclusively thought experiment, which, unlike the Turing test, assumed only a logical, not an actual possibility, acquired the features of an obvious absurdity, which gave an extra argument to those authors who insisted on the fundamental impossibility of the Chinese room experiment. That way, R. Penrose, who generally agreed with Searle's logic and considered the argument of the Chinese room considerably strong, admitted nevertheless that the program for the Chinese room is still too complicated for a person to make up (Penrose: 1989). Other authors further strengthened this thesis, claiming that no program is able to deceive a native of one or another language, say, Chinese, provided that he asks the machine the "right questions" (so-called "subcognitive" ones - for instance specially designed neologisms the meaning of which can only be perceived by a native speaker, questions about sensations, body control (what will happen if you put your hands together, etc.), contextual questions (what does a cloud look more like - the letter “x” or the letter “o”) and other similar ones (French: 2000, 460). To this we can add that it is not clear how a program, the essence of which is the manipulation of symbols separated from the content, is capable of answering even a simplest question such as "what is the time?"

It can certainly be said say that the latter objection is directed, first of all, against the Turing test, since Searle's thesis can be given in a different wording: "even if the machines passed the Turing test ideally, they would still comprehend neither the questions addressed to them nor their own answers." However, this objection touches upon another directly addressed to Searle's main thesis statement, according to which computer programs deal merely with syntax without touching upon semantics: even a purely syntactical program, suppose such one was feasible, would be extremely time-consuming to create and not very practical, therefore it would not correspond to the goals which artificial intelligence was created for - saving time and effort when performing certain mental operations. The algorithms of computer chess programs can serve an illustration here: if they remained at the syntax level, then in response to the opponent's move they would make one or another move in accordance with the rules, not caring about the strategy, or, which is less realistic, would recollect all the chess games existing in their memory in search of the appropriate position and the best move in it. However, the effectively playing programs give all the pieces involved in the game numerical values, i.e. "semantics" that takes into account the initial value of the piece as well as its position on the board and tend to outweigh the value of their pieces above the opponent's pieces.

Accordingly, many authors, who reacted critically to Searle's thesis, emphasized first of all the fact that Searle had separated syntax from semantics too radically whereas their relationship is much more complicated (Singh: 2014). Also, some insisted that syntax and semantics are simply inseparable (for instance the phrase "Maggie's Farther Leonard" is a type of semantic and simultaneously syntactic connection) (Boden: 1988, 263). Others added that syntax itself may not be sufficient for semantics but syntactic operations are capable of creating semantics (a syntactically correct combination of concepts is capable of creating new meanings: "today is Monday", etc.) (Speck: 2004). D. Chalmers, agreeing with Searle that a symbol itself is not capable of representing internal properties of an object, adds that it can transmit them through a system of relationships with other symbols (Chalmers: 1992, 30).

But if the negative part of Searle's thesis is built on a simplistically radical from the point of view of the majority of his opponent's separation of syntax and semantics, then its positive part stated just as unreasonable from their point of view convergence of thinking and its biological grounding (Rapaport: 1986). Many authors were also bewildered by Searle's interpretation of weak AI thesis as recognition of the computer's ability to simulate thinking in the same sense in which weather phenomena, such as a hurricane, can be simulated. Such comparison was unanimously considered unsuccessful since a computer performing, for instance, a counting operation, does not "simulate" a corresponding operation performed by the human brain, in the sense of simulation a meteorological phenomenon which cannot be confused with the phenomenon itself - it performs exact same operation (Rapaport: 1986, 15).

The listing of various positions from which Searle's thesis has been criticized can be continued, involving for example works which consider the substitution of the thesis contained in it (Maguire, aguire, Moser: 2015), other logical and gnosioligical mistakes (Damper: 2006). But even without this it is clear that in the overall balance of positive and negative evaluations of Searles's thesis the latter undoubtedly dominate, which seems to give every reason to believe that the thesis was ultimately rejected by science and suffered complete bankruptcy (Preston, Bishop: 2002, 46). Even the recognition by some authors of the "historical merit" of Searle's thesis has little effect on this final assessment: for example, V. Rapaport wrote that computer scientists should be grateful to Searle because he forced them to think about the grounding of their own theories (Rapaport: 1986, 18). The attempts to find a "limited use" of Searle's thesis, seeing in it an argument against computational theories of consciousness seem to have little effect on it as well.

Searle's Defence

Considering all this criticism, persistence with which Searle continues to defend his thesis looks strange. This persistence can as an option be explained by the scientist's predilection to the position he once expressed, human pride, and so on. But the matter does not seem to be just it. Perhaps the paradox here is that it is exactly the format of the Chinese room experiment, which is oriented to the Turing test, is the main obstacle for a convincing substantiation of the thesis illustrated by the experiment. To confirm this idea let us turn to Searle's "Rationality in Action" (2001) summarizing and continuing reflections which were the subject of his earlier works ("An Essay in the Philosophy of Mind" (1983), "Minds, Brains, and Science "(1984), "The Construction of Social Reality" (1995), and others). Although the experiment named the Chinese Room is not mentioned in the book of 2001, it can shed light on the hidden gnoseological grounding of the idea of the experiment, namely Searle's philosophy of consciousness.

But before talking about this we should return for a moment to the Turing test which was a reference point for the Searle's thought experiment, namely, the question of evaluating its possible results. The issue here is the inevitable considerable range of answers to many types of questions (and these are exactly those questions that do not imply the only "correct" answer and which are most suitable for identifying the "humanbeingness" of the respondent). Even obviously paradoxical and "stupid" answers to such questions will not mean that the respondent is not a human. What does this diversity point? In its most general terms it points that the procedures of thinking based on the whole set of life experience of an individual and not on the sum of rules learned by them cannot be formalized.

This can be said in a different way: the judgment expressed by an individual is always the result of a choice from a variety of alternative options and this choice cannot be completely determined. Between the grounding for choice and the choice itself there is what Searle in his book calls the "the gap". As Searle writes at the end of his book "The gap is a feature of human consciousness, and in this sense, my book is about consciousness". If, at the beginning of his work, Searle speaks of a conscious action, such as voting, for example, producing analytics of the "breaks" preceding and accompanying such action, he then expands the concept of conscious action, extending it, in principle, to any statement (saying "it is raining" I already assume responsibility for a certain state of affairs). At the same time following Sartre, whom however he does not refer to, Searle insists that "inside the gap" on one side of which there is an intention and on the other one there is an action, "there is nothingness".

But is "consciousness" synonymous with "conscious action" taken in the abovementioned expanded understanding? Not at all, says Searle: "We can make proof-theoretical or syntactical models, where the model exactly mirrors the substantive or contentful processes of actual human reasoning. And of course, as we all know, you can do a lot with the models. If you get the syntax right, then you can plug in the semantics at the beginning and it will go along for a free ride, and you get the right semantics out at the end because you have the right syntactical transformations" (Searle: 2001, 21). Such transformations are also conscious operations and are quite accessible to machines; moreover, such transformations according to the rules are carried out much better by machines rather than people and this is exactly the kind of tasks they were created for. It can also be noted that in the above statement Searle does not oppose syntax and semantics: according to the central thesis followed in the book "Rationality in Action" the gap which defines the essence of human thinking, since it does not refer to the abstract system of rules but to the world around the person, it is not a gap between meaning and symbol, neither it is a gap between a symbol and a symbol but it is a gap between meaning and meaning. Orientation from one meaning to another, overcoming (or not overcoming) separating their nothingness of the gap - is what Searle calls the "intentionality" of human consciousness.

As if continuing discussions on possible modifications of the Chinese room, Searle suggests to imagine a robot with a brain similar to that of a human, a motor apparatus, perception, memory. If we further place into this robot convictions, inclinations, speech and the ability to act on the basis of its thoughts, Searle asks whether this all will be enough for the robot to be able to act rationally. His answer is that if everything placed in this robot is a combination of algorithms, it will not, because: "the alleged ideal of a perfectly rational machine, the computer, is not an example of rationality at all, because a computer is outside the scope of rationality altogether. A computer is neither rational nor irrational, because its behavior is entirely determined by its program and the structure of its hardware" (Searle: 2001, 66). Searle is convinced that even the possibility of obtaining unpredictable results inserted in the program will be just an imitation of freedom and not the freedom that an individual performs when making decisions. Searle agrees that "for such a machine, consciousness might exist, but it would play no causal or explanatory role in the behavior of the system (Searle: 2001, 295)

Rationality as Action

At the same time, Searle is convinced that human consciousness also has its "hardware", being a function of a human body, and in a narrower sense - the activity of a human brain, which on its physiological part is strictly determined. How can rationality in action with its inherent freedom of choice can be combined with the determinism of the processes providing this activity? Using the hypothesis of successive discrete states, Searle responds that when, at any particular time, for instance t1, "the total conscious state of the brain, including volitional consciousness, is entirely determined by the behavior of the relevant microelements (Searle: 2001, 294); the same can be said about the subsequent moments t2 and t3, however, "the state of the brain at t1 is not causally sufficient to determine the state of the brain at t2 and t3 (Searle: 2001, 294). And "the move from the state at t1 to the state at t2 and t3 can be explained only by features of the whole system, specifically by the operation of the conscious self" (Searle: 2001, 294). Such scheme that likens a possible future theory of consciousness to quantum theory (Searle himself speaks of this analogy) looks, as Searle admits, completely speculative, while adding that rejecting such model we unwillingly tend towards this or that deterministic version of an explanation of human consciousness, assuming that psychological indeterminism coexists with neurobiological determinism. If that thesis is true, free rational life is entirely an illusion" (Searle: 2001, 298). Searle finishes his book with the conclusion that "conscious rationality is supposed to be a causal mechanism that proceeds causally, though not on the basis of antecedently sufficient causal conditions. Indeed, on some accounts, one of the functions of the cell is to overcome the instability of the quantum indeterminacy at subcellular levels" (Searle: 2001, 298).

CONCLUSION

It is clear that this conclusion itself is far from certainty and, moreover, it does not claim the status of the "last word" in the theory of consciousness, which Searle himself points as well, but it allows to take a slightly different look at the Chinese room experiment suggested by Searle, namely the so often criticized separation of syntax and semantics in the framework of this experiment. Without being afraid to be accused of tautology, we will try to introduce the following additional division into "syntactic semantics" and "semantic semantics". Their difference is clarified by the following example: I am playing chess and choosing between two variants of the next move, which seem to me equally strong. Here, the options for continuing the game are always determined, ultimately, by the syntactic rules and are subject to them. I am a mediocre chess player and almost any computer program will beat me because it better matches the syntax of the game with a specific position on the board; but the move that is made in the game is not just syntax - it is "syntactic semantics". Now, suppose I choose to vote for or against a proposal or a person. Will the machine which contains all the possible information on the subject under discussion have any advantage over me? I do not think so. The discussion procedure, of course, also has its own "syntax", but in this case the semantics is not derived from it and therefore gains a completely different form: it is "semantic semantics". It seems that the dichotomy suggested by Searle "human thinking / computer thinking" lies precisely here, and this, as it might be supposed, is the reason why according to one of the authors the so many times refuted "Chinese room argument seems alive and kicking" (Mooney: 1997). Actually, Searle does not deny the theoretical possibility of providing the robot with a complete analogue of the "neurophysiological uncertainty" inherent to a human brain but the general direction of his thoughts suggests that even such robot will lack a similar human body - with its own personal history and changing social environment.

BIBLIOGRAPHY

BODEN, M.A. (1988). Escaping from the Chinese Room. Computer Models of Mind. Cambridge University Press, 253-266.

CHALMERS, D.J. (1992). Subsymbolic Computation and the Chinese Room. The Symbolic and Connectionist Paradigms. Lawrence Erlbaum Associates, 25–48.

DAMPER, R.I. (2006). The logic of Searle’s Chinese room argument. Minds and Machines, 16(2), 163–183.

FRENCH, R.M. (2000). The Chinese Room: Just Say “No!”. Proceedings of the 22nd Annual Cognitive Science Society Conference. NJ. LEA, 657-662.

MAGUIRE, R., MAGUIRE, PH., MOSER, PH. (2015). A clarification on Turing’s test and its implications for machine intelligence. Proceedings of the 11th International Conference on Cognitive Science, 318-323.

MOONEY, V.J. Iii. (1997). Searle's Chinese Room and its Aftermath: Ph.D. candidate. Stanford: Computer Systems Laboratory Stanford University. Retrieved from: https://www.cs.rit.edu/~mpv/course/ai/chinese-room.pdf

PENROSE, R. (1989). The Emperor’s New Mind: Concerning Computers, Minds, and the Laws of Physics. Oxford University Press, 480.

PRESTON, J., BISHOP, M. (2002). Views into the Chinese Room: New Essays on Searle and Artificial Intelligence. New York. Oxford University Press, 410.

RAPAPORT, W.J. (1986). Philosophy, Artificial Intelligence, and the Chinese-Room Argument. Abacus, 3(4), 7-18.

SAYGIN, A.P., CICEKLI, I. (2000). Turing Test: 50 Years Later. Minds and Machines, 10(4), 463-518.

SCHANK, R.C., ABELSON, R.P. (1977). Scripts, plans, goals, and understanding. Hillsdale, N.J.: Lawrence Erlbaum Press, 248.

SEARLE, J.R. (1980). Minds, brains, and programs. Behavioral and Brain Sciences, 3(3), 417-457.

SEARLE, J.R. (1990). Is the Brain's Mind a Computer Program? Scientific American, 262(1), 26 - 31.

SEARLE, J.R. (2001). Rationality in Action. Cambridge, Massachusetts, London: A Bradford Book, 319.

SINGH, A. (2014). What is really in the Chinese Room? Trinity College Dublin Category. http://www.undergraduatelibrary.org/2014/philosophy-theology/what-really-chinese-room

SPECK, E VAN DER. (July 2, 2004). Searle’s Chinese Room Argument. Retrieved from: https://spekmen.home.xs4all.nl/remote/ChineseRoom.pdf

TURING, A.M. (1950). Computing Machinery and Intelligence. Mind, 49, 433-460.

BIODATA

Maxim Aleksandrovich MONIN: Candidate of Philosophy, Associate Professor at Department of Humanities, I.M. Sechenov First Moscow State Medical University (Sechenov University). Previous job - Department of Cultural Studies of the St. Tikhon’s Orthodox University (2006 - 2017). Graduated from Lomonosov Moscow State University, Faculty of Philosophy. Interests: history and theory of culture, history of philosophy, modern philosophy, philosophy of science.

Vera Albertovna TEREKHOVA: Candidate of Philosophy, Associate Professor at Department of Humanities, I.M. Sechenov First Moscow State Medical University (Sechenov University). Graduated from Lomonosov Moscow State University, Faculty of Philosophy. Interests: history of philosophy, philosophy of culture, philosophy of religion, philosophy of science.

Elena Vladimirovna LEDENEVA: Candidate of Philosophy, Associate Professor at Department of Humanities, I.M. Sechenov First Moscow State Medical University (Sechenov University). Graduated from Lomonosov Moscow State University, Faculty of Philosophy. Interests: history of philosophy, philosophy of culture, philosophical anthropology, philosophy of science.