The Journal of Mind and Behavior
Summer and Autumn 2023, Volume 44, Numbers 3 and 4
Pages 147–156
ISSN 0271–0137

Is a Conscious Robot a Scientific Hypothesis or Just a Faith?

Sam S. Rakover

Haifa University

I try to show that the hypothesis that a very sophisticated robot may develop consciousness and understanding is not a scientific hypothesis. First I argue that a robot that perfectly imitates human behavior will not necessarily be endowed with consciousness. I then argue, based on a thought experiment that I call the "robotic-mom," that the hypothesis about a conscious robot is just a matter of faith.

Keywords: AI, consciousness, robots

The goal of "strong AI" is to create a human-like machine (a computer or a robot) that, like human beings, will behave intelligently and that will have other human characteristics, especially consciousness and understanding. In contrast, the goal of "weak AI" is to build a machine that may appear human-like. In discussing strong AI, Searle (1980) writes as follows: "… 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 cognitive states" (p. 417). The purpose of the present article is to show that strong AI is nothing but faith (see reviews, e.g., Bringsjord and Govindarajulu, 2022; Cole, 2020; Hauser, 2023; Rakover, 1999).

Searle's (1980) famous thought experiment, the Chinese Room, concludes that, although a highly sophisticated and complex computer might be able to pass the Turing test (according to which, one cannot differentiate between intelligent behavior of a human and a computer), it would be incapable of developing consciousness and understanding. After a brief discussion of this thought experiment, I will focus on two additional arguments that may be taken to support the conclusion of the Chinese Room argument. First, I will defend the idea that a robot that successfully imitates human behavior is nothing more than one

I would like to express my gratitude to the journal editor, Raymond Russ, for his insightful and useful comments that improved greatly the paper. Correspondence should be addressed to Sam S. Rakover, Ph. D. Department of Psychology, Haifa University, Haifa, Israel 31905. Email: rakover@psy.haifa.ac.il

possible mechanical system of many other different mechanical systems. However, from this ability to imitate it does not necessarily follow that a sophisticated robot develops human consciousness and understanding. Second, I will suggest a thought experiment, the robotic-mom, which will show that the hypothesis according to which a highly complex and sophisticated robot will have consciousness and understanding is not a scientific hypothesis but a mere faith.

In brief, Searle's (1980) Chinese Room thought experiment is structured as follows. Chinese scientists have developed a complex and very sophisticated computer, capable of receiving questions and producing intelligent answers in Chinese. Searle, who does not understand Chinese, decides to take upon himself the role of that computer. On one side of the Chinese Room, Searle receives an input written in Chinese and, with the help of an English instruction book that guides him in manipulating the Chinese characters, he produces an output written in Chinese. Chinese scholars, who are outside this room, judge Searle's answers to be correct and highly intelligent. However, at the end of the experiment, when Searle leaves the Chinese Room, he announces that he did not understand a word in Chinese. Because Searle, who has consciousness and understanding like any other human being, announces that he does not understand Chinese, it is clear that the sophisticated computer does not understand Chinese either, even though it is programmed to handle inputs in this language in the most successful way. Although this thought experiment has provoked many ongoing debates, it would be reasonable to suggest that this experiment is considered, to this day, to be one of the strongest arguments against strong AI (e.g., Bringsjord and Govindarajulu, 2022; Cole, 2020).

Multi-Functionality

I suggest conceiving the term "multi-functionality" as a general concept that refers to several phenomena as follows: (1) many roads lead to Rome, that is, a goal can be achieved through many different means; (2) many different functions can be fitted to a given set of empirical observations; (3) it is possible to offer many different interpretations of a given written passage; (4) it is possible to offer many different realizations for a given function. What all these cases have in common is that there are many alternative ways of describing or carrying out a certain goal or a function (e.g., Kim, 1996; Nola and Sankey, 2007). In light of this, we may propose that a robot, no matter how sophisticated it may be, is just another example of multi-functionality. We can conceive of a robot (or a computer) as one among other ways of imitating human behavior. Now, given that many mechanical systems succeed in imitating human behavior, can one expect them all to develop consciousness and understanding? This problem requires two comments on the subject of analogy and on that of fear.

Analogy. By now almost everyone knows that robots can perform hundreds of human-like responses and actions: robots can walk, run, pick up things and

put them in their proper locations with great accuracy, cut wood and metals with great precision, they can overcome obstacles, help in high-technologies, they answer questions logically, perform medical diagnoses, beat masters in chess and go, they can learn to speak, write poems, stories, music, paint, and so forth.

However, the fact that robots can imitate human behaviors so well does not necessarily imply that they can also develop consciousness and understanding. This is so because the hypothesis that a robot might develop consciousness and understanding can be seen as based on an analogical argument. If the target system (the robot) is similar in many details to the source system (the human), then one may suggest that consciousness and understanding will also develop in a robot as these qualities have developed in humans. Bartha (2022) writes: "The more similarities (between two domains), the stronger the analogy" (subsection 3.1). Thus, (a) the greater the similarities we find in intelligent behavior between a human and a robot, the stronger the analogy and, therefore, (b) the greater the support of the hypothesis that eventually a robot will come to possess consciousness and understanding. The problem is that (b) does not necessarily follow from (a) because an analogical argument is inductive and ampliative in nature. For example, there are very many similarities between a man and a donkey. However, the donkey still cannot respond correctly to an arithmetic question (even though one may hypothesize that the donkey simply refuses to answer this question — after all, donkeys are known to be very stubborn). So, in light of the multi-functionality of intelligent human behavior, one may suggest that a robot's behavior is just another way to describe and perform human-like behavior. And, since the above hypothesis about a conscious robot can be understood as based on an analogical argument, it does not necessarily follow that a robot will develop consciousness and understanding.

Fear. Because robots can do many things that humans do, the following observation becomes salient. While humans are not afraid of robots helping in industries, they are filled with fear when robots succeed in imitating human behavior which indicates cognitive intelligent abilities. The fear pertains to the conjecture that at a certain stage in robot evolution, a critical stage will be reached in which robots will develop consciousness and understanding, which may even surpass human intelligence. This fear is expressed in science fiction (such as Asimov's *I, Robot*) and apocalyptic films involving robots with consciousness and understanding waging war on human beings (for example, Schwarzenegger's movie *Terminator 2: Judgment Day*). Furthermore, AI researchers and newspaper articles warned the world of an AI apocalyptic destruction of civilization (for example, see *New York Post*, July 17, 2023, "5 Terrifying Stories that Warn of an AI Apocalypse"). Given this, one may ask: Is this fear based on a scientific hypothesis, according to which very complicated and sophisticated robots will develop consciousness and understanding, or is this fear nothing more than a matter of

a faith that indeed sophisticated robots will develop consciousness and destroy humanity? The next section argues for the latter option.

The Robotic-Mom Thought Experiment

The multi-functionality of intelligent human behavior illustrates that a sophisticated robot may well imitate human behavior even though the robot does not necessarily develop consciousness and understanding. The purpose of the current thought experiment, the robotic-mom, is to reinforce the claim that an advanced and sophisticated robot will not develop consciousness. This thought experiment does this by supporting the following suggestion: anyone who holds the hypothesis that an advanced and sophisticated robot may develop consciousness is in effect committed not to a scientific hypothesis but only to a certain faith. A scientific hypothesis is a hypothesis that fulfills the methodological requirement for rigorous empirical testing (e.g., Keas, 2018; Nola and Sankey, 2007; Popper, 1959; Rakover, 1990). According to Popper (1959), it must be falsifiable; it must admit itself to the possibility of being refuted. For example, in psychology, the procedure of conducting experiments requires that from every hypothesis it is possible to derive a prediction that admits itself to empirical testing that either confirms or refutes the hypothesis (e.g., Chow, 1987; Neal and Liebert, 1986; Rakover, 1990, 2003).

The robotic-mom thought experiment allows for a decisive decision between two contrary hypotheses:

(a) H(c*): at some critical stage of advancement and sophistication a robot will develop consciousness and understanding (where c* designates consciousness). Several researchers believe that sophisticated robots, or computers, will develop consciousness (for review and discussion see e.g., Buttazzo, 2001; Chella et al., 2019; Koch, 2018; Reggia, 2013). Furthermore, several researchers have suggested the term "singularity" to denote "… the future point at which artificial intelligence exceeds human intelligence, whereupon immediately thereafter (as the story goes) the machines make themselves rapidly smarter and smarter and smarter, reaching a superhuman level of intelligence that, stuck as we are in the mud of our limited mentation, we can't fathom" (see Bringsjord and Govindarajulu, 2022, section 9).

(b) H(nc*): at no stage of advancement and sophistication will a robot develop consciousness and understanding.

For the robotic-mom thought experiment to determine between these two hypotheses, it must meet the following requirements.

(1) Since consciousness is a state of mind characteristic of each individual that only that individual is capable of experiencing (following Nagel, 1974), the

robotic-mom thought experiment has to allow for an intuitively strong connection between consciousness and a particular public behavior.

(2) This particular public response must be decisive so that its appearance would support one hypothesis and the appearance of an opposite response would refute it, i.e., it has to confirm or refute the above two hypotheses: H(c*) or H(nc*).

As will be seen, the robotic-mom thought experiment meets these two requirements nicely. The thought experiment is as follows. Scientists have created a robotic-mom that imitates with absolute precision the behavior of a human-mother taking care of her newborn baby (human-baby). When the nurse brings the human-baby to the human-mom, she smiles warmly, embraces her baby, cradles him in her arms with love, and feeds him. This behavior, which is normal behavior expected from a mother after giving birth, reflects the human-mother's state of mind. The robotic-mom exhibits exactly the same behavior when the nurse brings her the human-baby. Suppose that the nurse accidentally brings a robotic-baby to the human-mom. The human-mom's reaction is immediate, clear, and expected: she rejects the robotic-baby. The critical question of this thought experiment is: What will the robotic-mom do when the nurse brings the robotic-baby to her? Two simple and opposing responses are possible.

Response (A). According to H(nc*), the robotic-mom will mimic the human-mom exactly, i.e., she will reject the robotic-Baby.

Response (B). According to H(c*) the robotic-mom will *not* mimic the human-mom's behavior and will *not* reject the robotic-baby. On the contrary, she will smile at the robotic-baby, embrace him, and feed him. *Why*? Because *if* she has developed consciousness, *then* when she sees the robotic-baby she will understand that this baby is her own kind and is her real baby. Furthermore, if the human-baby were brought to her now, she would reject him. In other words, the robotic-mom will realize that a human-baby cannot be her own, because he is made of a different material than she is made of, while the robotic-baby is made of the same material she is made of.[1]

Although there is no empirical evidence concerning these two hypothetical responses, the scientific status of the two hypotheses, H(c*) and H(nc*), can be evaluated. It is clear that H(nc*) meets the accepted and most important requirement of a scientific hypothesis, refutability. Accordingly, if the robotic-mom does *not* reject the robotic-baby, but accepts it warmly (Response B), H(nc*) will be refuted. This result should be accepted at any stage of the robot's development; so long as the robotic-mom does not reject the robotic-baby, H(nc*) will be falsified. In this respect, the current experiment is crucial for H(nc*).

---

[1] Rakover (2021) proposed that consciousness is a necessary condition for understanding. Human beings without consciousness cannot understand the situation they are in and what they have to do. Furthermore, when humans lose consciousness they cannot even stand on their feet.

This situation does not hold for H(c*). According to this hypothesis, if the robotic-mom develops consciousness and understanding, she will not elicit response (A), but will elicit response (B), i.e., she will accept the robotic-baby. This response testifies that she has indeed developed consciousness and understanding that she is not acting merely as a machine that imitates the behavior of a human-mom, but that she consciously understands that the robotic-baby is her own. But what happens if the robotic-mom rejects the robotic-baby? Is H(c*) refuted? The answer is no!

If the robotic-mom rejects the robotic-baby, this observation would not count as a refutation of H(c*). Why? Because it is possible that the robotic-mom has not yet reached the critical stage of advancement and sophistication that allows for the creation of consciousness and understanding. The problem with this argument, which on its face seems attractive and rational, is that it makes H(c*) an ad hoc hypothesis. The argument arising from H(c*), which appeals to some critical stage of advancement and sophistication, can be seen as an argument whose only function is to save H(c*) from falsification; it can be invoked whenever H(c*) has been refuted. One may use this ad hoc argument whenever the robotic-mom, no matter how advanced and sophisticated it is, fails to develop consciousness and understanding, which will manifest in its rejecting the robotic-baby. Therefore, H(c*) is immune to empirical testing, to the possibility of refutation. Thus, it can be seen as a mere faith, not as a scientific hypothesis.

In conclusion, an analysis of the robotic-mom thought experiment shows that while H(nc*) has the status of a scientific hypothesis, H(c*) is nothing more than faith: the belief that one day in the future (near or far) the sophistication and development of robots will reach the level that they will develop consciousness and understanding.

Discussion

The criterion of refutation is one of several criteria for evaluating a scientific theory, alongside simplicity and fruitfulness (e.g., Keas, 2018; Nola and Sankey, 2007; Popper, 1959, 1963; Rakover, 1990). Popper (1959) conceives of falsification as central to his philosophy of science: a demarcation line between scientific and non-scientific theories. This criterion has been severely criticized, but in the context of the present article, I will only briefly discuss the criticism that arises from Duhem's problem (e.g., Duhem, 1996; Harding, 1976; Rakover, 2003). Duhem (1996) suggests "that an experiment in physics can never condemn an isolated hypothesis, but only a whole theoretical group" (p. 8). This is so because the tested prediction is derived from a theory, T, along with a set of auxiliary hypotheses and background theories. Thus, when the prediction is refuted (the prediction does not fit the experimental finding) the whole group [T and auxiliary hypotheses and background theories] is falsified and one cannot discern where the error lies.

Although this criticism is logically valid, Rakover (2003) suggested a practical method, in psychology, by which one may discern at which element of the group [T and auxiliary hypotheses and background theories] the experimental result is directed.

Given that scientists do not tend to abandon a falsified theory, Popper (1959, 1963) suggested that T can be saved from refutation with the help of an auxiliary hypothesis that allows for the generation of a new prediction. If the auxiliary hypothesis does not allow such a prediction, it is nothing more than an ad hoc hypothesis. As an example of this kind of method, he offered the discovery of the planet Neptune. Astronomers observed that the orbit of Uranus deviates from the prediction derived from the Newtonian mechanics. To save this theory, they proposed the existence of a planet, located at a certain place, that would influence the orbit of Uranus — and, so, Neptune was discovered.

H(c*) does not behave like an auxiliary hypothesis that generates a new testable prediction; it is more similar to an ad hoc hypothesis that functions only to salvage itself from refutation. The main reason is that one can always appeal to the critical stage of a robot's advancement and sophistication to reject the recalcitrant observation (experimental result) when an advanced robot fails to develop consciousness and understanding. Further, no one knows how exactly to define theoretically and practically this critical stage. As a result, H(c*) is immune to the process of falsification and becomes a matter of mere faith that, one day in the future, robots will pass the critical stage and be endowed with consciousness and understanding.

Several researchers have suggested that a robot's level of complexity is the key for generating consciousness and understanding (e.g., Churchland and Churchland, 1990; Dennett, 1991; Kim, 1966). For example, Dennett (1991, p. 440) writes (in the context of the Chinese Room argument):

> Complexity does matter. If it didn't, there would be a much shorter argument against strong AI; "Hey, look at this hand calculator. It doesn't understand Chinese, and any conceivable computer is just a giant hand calculator, so no computer could understand Chinese. Q.E.D." When we factor in the complexity, as we must, we really have to factor it in — and not just pretend to factor it in. That is hard to do, but until we do, any intuitions we have about what is "obviously" not present are not to be trusted.

Of course, the problem is that we do not know what structures and levels of complexity are needed to create consciousness and understanding.

So, I can suggest that as long as H(nc*) has not been refuted it is appropriate to hold it as a good and effective scientific hypothesis. In contrast, hypothesis H(c*) is nothing more than faith. However, as we will see, it still has some benefits.

The interesting point here is that H(c*) may energize the development of sophisticated robots even more than H(nc*). The reason for this lies in the

researcher's motivation. H(nc*) may stimulate the researchers to develop sophisticated robots to show that they do not develop consciousness and understanding. But H(c*) may stimulate the researchers to pursue one of the most important goals in the world, to solve the great mystery of consciousness, the mind–body problem, and the relation between consciousness and the brain. (Here I ignore all the other important personal and social goals researchers may have, e.g., money, prestige, developing an optimal robot, improving quality of life, etc., goals that, in practice, may greatly stimulate researchers to work hard in the field of robotics and computers.)

In light of the above, I propose that H(c*) is nothing but faith. And if this is so, it would be appropriate to suggest that complex and sophisticated robots should be treated as nothing more than efficient tools, with great potential to benefit humanity. The problem, of course, is that one has to learn how to handle this powerful tool, and this seems to require profound changes in the educational and judicial systems.

References

Bartha, P. (2022). Analogy and analogical reasoning. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2022 edition). Retrieved from https://plato.stanford.edu/archives/sum2022/entries/reasoning-analogy/

Bringsjord, S., and Govindarajulu, N. S. (2022). Artificial intelligence. In E. N. Zalta and U. Nodelman, (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2022 edition). Retrieved from https://plato.stanford.edu/archives/fall2022/entries/artificial-intelligence/

Buttazzo, G. (2001). Artificial consciousness: Utopia or real possibility? *Computer*, *34*, 24–30.

Chella, A., Cangelosi, A., Metta, G. and Bringsjord, S. (2019). Editorial: Consciousness in humanoid robots. *Frontiers in Robotics and AI*, *6*, Article 17.

Chow, S. L. (1987). *Experimental psychology: Rationale, procedures, and issues.* Calgary, Allberta: Detseling Enterprises Ltd.

Churchland, B. J., and Churchland, P. S. (1990). Could a machine think? *Scientific American*, *262*, 26–31.

Cole, D. (2020). The Chinese room argument. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2020 edition). Retrieved from https://plato.stanford.edu/archives/win2020/entries/chinese-room/

Dennett, D. C. (1991). *Consciousness explained*. Boston: Little, Brown and Co.

Duhem, P. (1996). *Essays in the history and philosophy of science.* Cambridge: Hackett Publishing Company.

Harding, S. G. (1976). *Can theories be refuted? Essays on the Duhem–Quine thesis.* Dordrecht, Holland: D. Reidel Publishing.

Hauser, L. (2023). Artificial intelligence. *The Internet encyclopedia of philosophy.* ISSN 2161–0002.

Keas, M. N. (2018) Systematizing the theoretical virtues. *Synthese*, *195*, 2761–2793

Kim, J. (1996). *Philosophy of mind*. Boulder: Westview Press

Koch, C. (2018). What is consciousness? *Nature*, *557*, S9–S12

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, *83*, 435–450.

Neale, J. M., and Liebert, R. M. (1986). *Science and behavior: An introduction to methods of research.* Prentice–Hall.

Nola, R., and Sankey, H. (2007). *Theories* of *method: An introduction.* Stocksfield, England: Acumen.

Popper, K. (1959). *The logic of scientific discovery.* London: Hutchinson.

Popper, K. (1963). *Conjectures and refutations.* London: Routledge and Kegan Paul.

Rakover, S. S. (1990). *Metapsychology: Missing links in behavior, mind and science.* Paragon/Solomon.

Rakover, S. S. (1999). The computer that simulated John Searle in the Chinese Room. *New Ideas in Psychology*, *17*, 55–66.
Rakover, S. S. (2003). Experimental psychology and Duhem's problem. *Journal for the Theory of Social Behaviour*, *33*, 45–66.
Rakover, S. S. (2021). The two factor theory of understanding (TFTU): Consciousness and procedures. *Journal of Mind and Behavior*, *42*, 347–370.
Reggia, J. (2013). The rise of machine consciousness: Studying consciousness with computational models. *Neural Networks*, *44*, 112–131.
Searle, J. R. (1980). Minds, brains and programs. *Behavioral and Brain Sciences*, *3*, 417–457.