**Chapter (8): Could a sophisticated robot develop consciousness?**

**Pat and Rick drank a lot of beer in ‘our bar’. They were frustrated from their research in the elusive phenomenon of consciousness. Pat quickly took a sip from the bottle and said, "Listen Rick, listen, listen, I have another big idea to figure out why we can't grasp the elusive creature in ‘our bar’. Here's what I think. This creature is some kind of robot, doing everything we do but we can't catch this! So what do you say, Rick?" Rick pulled the corners of his mouth down a bit and said, "Just a small question Pat, can we buy him a beer?"**

The attempt to answer the question of the chapter is only one of many goals and problems that the new scientific field of computer science deals with. Particularly, this question is a significant issue of the specific field called: “the philosophy of artificial consciousness (AC)” (e.g., Chrisley, 2008: Müller, 2024; Russell and Norvig 2003). I am not going to treat all their issues here (for an overview see McCorduck, 2004). I will concentrate on the attempt to offer an answer to the chapter's problem. That is to say, I will address the problem that Searle (1980) calls "strong AI", the question of whether an appropriate computer program is indeed a mind.

Given this, I would like to precede and present here at the beginning of the present chapter the main conclusion arising from the debate about the chapter’s question. The answer to this question is that it is ultimately a matter of one’s own conviction, belief – one’s own degree of persuasion in the possibility of a robot developing consciousness (CΨ). An important reason why I move the level of discussion from that of proof, confirmation/refutation, to the level of persuasion and belief, is as follows. It is not possible to be perfectly sure, to be in one hundred percent conviction, that the other person feels exactly what we feel under the same conditions (this is actually the problem of the other mind that does not have a solution accepted by everyone, see Avramides, 2023).

There are those who may believe that very sophisticated robots (computers, software) have indeed developed or will develop eventually CΨ, such as, for example, Blake Lamoine, a Google engineer, who claims that the Google "Lamda" software has CΨ (e.g., Blackmore, 2013; Vallance, 2022. There are several famous researchers who supported the idea of ​​human-like AI, such as, Turing, Simon, Newell, Minsky, McCarthy, McDermott, Chalmers, Tononi, and Kurzweil (e.g., McDermott, 2007). Other researchers, like me, believe that sophisticated robots have not yet developed CΨ and the chances are very high that robots will not develop CΨ similar to that of humans and animals.

The possibility that highly complex and sophisticated robots will develop CΨ like us humans has evoked in humans a large spectrum of reactions, emotions, attitudes and different arguments related to our question. In the USA, a number of surveys of attitudes towards AI were conducted. Faverio and Tyson (2023) report that 52% of the participants were concerned about the impact of AI in their daily lives, and 60% felt uncomfortable with the introduction of AI into medical considerations. Elsey and Moss (2023) report that 51% of the participants were interested in a pause on AI research, and 67% think that AI will surpass human intelligence. Although these surveys contain other interesting information, I believe that the presented information can be interpreted as supporting the hypothesis that AI raises concern and anxiety. Why? Perhaps because of the following observation. 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. [In fact, there are quite a few cases where humans relate positively and kindly to humanoid robots (e.g., Blackmore, 2013).] The fear of sophisticated robots pertains to the conjecture that at a certain stage in robot evolution, a critical stage will be reached in which robots will develop CΨ 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 CΨ and understanding waging war on human beings (for example, Schwarzenegger’s movie [*Terminator 2: Judgment Day*](https://www.imdb.com/title/tt0103064/?ref_=ttls_li_tt)). Furthermore, AI researchers and newspaper articles warned the world of an AI apocalyptic destruction of civilization (e.g., see *New York Post*, July 17, 2023, "5 terrifying stories that warn of an AI apocalypse").

Why we fear that machines imitating human behavior will develop CΨ, but we have no fear of a vast collection of machines whose performance is much better than that of humans. For example, cars that transport humans over land, sea and air, devices that manage to cut metals with great precision, combines that harvest grain and collect them in convenient piles for their collection, devices that heat water to a desired temperature, etc.

Why don't we expect the heroine of Tolstoy's novel "Anna Karenina" to get out of the pages of the novel and actually commit suicide? Why are we not afraid that a chemical formula of dynamite, which is written on a piece of paper, will explode even if the paper happens to catch fire? Why are we not afraid that the cruel hero that is appearing in a horror movie will come out of the screen and start slaughtering the audience? However, we do fear that a certain robot (or a complex and sophisticated computer program) will develop CΨ and as a result will surpass man in its wisdom and malice! Why? After all, in all these and other similar cases, we are dealing with the most interesting creations of humans, entities that have no life of their own. What is the difference between Anna Karenina, a chemical formula, an evil hero in a horror movie, and a robot that imitates human behavior? All these (and others) are nothing more than the creations of a human being, the fruit of man's mind that testify to nothing more than the power of his/her imagination and thinking. There is no difference at all! In each and every case, a person invested the peak of his intellectual and imaginative power. And this is exactly how these should be treated – as human’s creations. One should not expect these creations themselves to be transformed into living beings with CΨ. (Humans are of course capable of creating other beings with CΨ. Everyone knows that - these are their children.)

In the light of the above, it seems to me that the interesting question here is not whether a sophisticated robot will develop CΨ, but the question is: to what do we attribute human qualities, consciousness, and why. The reasons are many be anchored to human's emotions or rationality.

From the emotional point of view, the old story of the Greek sculptor Pygmalion is enlightening. After sculpting the statue of the woman, Galatea, Pygmalion fell in love with that statue. According to legend, his love for the statue was so strong that Aphrodite, the goddess of love, turned Galatea the statue into a flesh and blood woman. That is, apparently man has a tendency to treat beings similar to him as living beings and in the case of Pygmalion to fall in love with the beautiful statue of Galatea.

From a rational point of view, it is possible, for example, to conceive of one’s mental-state from the functional approach, from its causal relations to stimulus, response, other mental states and the realizations of its function. Given this, it is possible to propose that CΨ can be understood in terms of its functions. Therefore, if a robot fulfills the CΨ-functions in the same way as a human, it would be justified to claim that the robot is endowed with CΨ. I will now briefly review a number of arguments supporting the possibility of a conscious-robot, move on to the counter-arguments, and end with some arguments, thought experiments and “experimental stories”, which I have composed over the years that I have contemplating about the subject, with the aim to convince of the impossibility of a conscious-robot.

*Machines with artificial consciousness?*

Chella and Manzotti (2011), who discuss the issue of a machine-CΨ, point out that the construction of machines of this type (with CΨ) does not depend entirely on a particular theoretical approach, such as, the computational approach to cognition and CΨ. They write (p. 640): "Luckily, artificial consciousness does not need to be committed to a particular theoretical view since it can appeal to real implementation that would hopefully overcome any theoretical limits." According to them, a machine-CΨ is between two extremes: only the brain produces CΨ, and CΨ is a functional system that can be implemented in multiple ways. They described and reviewed several robots and models that imitate different cognitive behaviors related to CΨ (see also Blackmore, 2013). Have these machines developed human-like CΨ? I don't think so. Similarly, (Chella and Manzotti (2011, p. 647) write: "Human beings (and perhaps animals) have intrinsic intentionality that gives meaning to their brain states. On the contrary, artificial systems like computers or stacks of cards do not have intrinsic intentionality."

*Arguments for the possibility of a conscious-robot*

(a) *Copying, replacing the brain*. Several researchers deal with the idea of a thought experiment, in which the brain cells are replaced with artificial units (e.g., silicon chips) in order to test whether CΨ is preserved in the artificial brain. I will discuss two, Kurzweil (1999) and then Chalmers (1996).

Kurzweil (1999) opens his book with a thought experiment that is a variation on the Greek thought experiment, called "The Ship of Theseus". The Greek myth describes the ship of Theseus after he killed the Minotaur with Ariadne's help. The Athenians kept this ship as a memorial and from time to time replaced very old beams with new ones, until the whole ship was renewed, the “renewed-ship”. Furthermore, the Greeks kept all the old beams and built form them the "rebuilt-ship". And here the following question arises, which of the two ships is that of Theseus, (a) the renewed-ship or (b) the rebuilt-ship? This question raises the following philosophical problem: does a being retain its identity even if all its parts are replaced (e.g., Blackbum, 2024).

Kurzweil (1999) proposes, among other variations, to scan a person called Jack and based on this scanning to build an identical copy of him. Will Jack's CΨ be preserved in the "Copied-Jack"? If CΨ depends solely on the format of the particles that make up Jack (consider his brain) on their functional relationships (according to the functionalist approach), then the answer is positive. However, if we take into account the material from which a human brain is constructed (Jack’s brain), we will offer a negative answer, because the material of the Copied-Jack is different from that of the original Jack. (Note that in the case of the Ship of Theseus, it is also possible to take into account the element of time and suggest that the rebuilt-ship is the original ship, because its old beams are the original beams of that ship, which have undergone an ageing process over time.)

One of Chalmers (1996) variations on the idea of copying, replacing the brain, is the Fading-Qualia argument. The main idea is to support the hypothesis that CΨ depends on the structure of the brain, on its functional organization. Chalmers considers two beings, X is Chalmers himself, who has a certain brain functional organization and CΨ, and X\* who has the same brain functional organization but does not have CΨ (since X\* brain consists of silicon chips). That is, X\* is Chalmers’ Zombie. Now one start replacing the neurons in the brain of X (Chalmers) with silicon chips. The question that arises here is this: does the CΨ of X remain as it is or does it change as X’s brain gets similar to X\* brain, which lacks CΨ. Chalmers puts forward arguments that show that CΨ cannot suddenly disappear completely or that CΨ is not fading gradually. Therefore, the following possibility is true: X’s CΨ remains intact; a conclusion that supports the idea that CΨ depends on the brain functional organization, and also artificial intelligence (AC).

This thought experiment raised a number of criticisms, some of which Chalmers addressed in his book. Here I would like to summarize my Contra-Zombie (CZ) argument (see Chapter 6) which seems to be relevant to the present thought experiment. Briefly, the CZ argument proposes: A Zombie is an *active* person. However, since a Zombie is a person without CΨ, who is *inactive* (or *dead*) then a Zombie is *inactive*. Thus, a Zombie is an *active* and *inactive* at the same time. This conclusion shows self-contradiction and therefore it contradicts Zombie conceivability – it makes no sense. Given this, the CZ argument undermines the foundations of the Fading-Qualia thought experiment, as a Zombie is inconceivable.

(b) *The computational theory of mind (CTM).* The major idea of the CTM approach, which forms the basis of cognitive science, is that the human brain/mind is a computational system, a kind of computer (e.g., McDermott, 2007; Milkowski, 2024; Rescorla, 2020;[Russell](https://en.wikipedia.org/wiki/Stuart_J._Russell) and [Norvig,](https://en.wikipedia.org/wiki/Peter_Norvig) 2003). Many researchers believe that cognitive processing have to be understood as computations on mental representations. One interpretation of this kind of processing is that physical symbols (as in a computer) are manipulated by using certain rules. In the light of this, it is possible to suggest that a robot, a computer, based on the appropriate software/hardware, may also develop CΨ, artificial CΨ, AC. For example, McDermott (2007) proposes, within the framework of the 'computational models of consciousness', the ideas that CΨ is a property of a system that is capable of creating a model that represents the system itself, and that a highly intelligent, complex and sophisticated computer will lead to the development of AC.

*Arguments against the possibility of a conscious-robot*

There are several arguments against the possibility of a conscious-robot (computer). I shall briefly review the prominent and the well-known ones. (It is worth emphasizing that the criticisms against the CTM has provoked a lively debate in the professional literature that I cannot detail here.)

(1) *Arguments based on Gӧdel’s incompleteness theorems.* In a very simplistic way, it can be proposed the following. Gӧdel's first theorem says that for every formal consistent system (S), which is free of contradictions and capable of handling arithmetic, there are statements that cannot be proven true or false in S, that is, S is an incomplete system. (Note that *consistency* means free of contradictions, and *completeness* means proof or disproof of any statement.) Gӧdel's second theorem says that a formal system that includes arithmetic cannot prove its own consistency. That is, to prove that a S does not contain contradictions, one need to appeal to another system outside of S. As an example of a mathematical hypothesis, which appears to be true (and supported by many examples) but which have not yet been proven true or false, is the Goldbach conjecture. According to this conjecture, every even number greater than 2 is the sum of two prime numbers (e.g., 16=5+11; 30=13+17).

Lucas (1961) and Penrose (1989, 1994) based their support for the argument that the human brain/mind is not like a computer mainly on Gӧdel's first theorem. Lucas's (1961) fundamental argument is as follows. Given (a) a formal machine, which produces proofs of truth, cannot handle a sentence called “Gӧdel sentence” (e.g., “This formula is unprovable-in-the-system”, Lucas, 1961, p. 112), and (b) a human brain/mind is capable of understanding the truth of this sentence, then the human brain/mind is not like a computer, it cannot be explained as a machine. Penrose (1989, 1994) suggested that mathematicians conceive intuitively and consciously the truth of a non-computable sentence. Thus, conscious comprehension is different from understanding through computation. Moreover, since Penrose believes that our brain/mind is explained by physical theory, he proposed that conscious comprehension is not based on computable classical physics, but on quantum theory (see chapter 2).

The arguments of Lucas and Penrose have provoked a large and lively debate in the professional literature (e.g., Megill, 2024; Selmer and Govindarajulu, 2024). For example, one of the counterarguments against Lucas's (1961) argument deals with the question of whether human cognitive system is consistent. If the answer is no, then everything can be proven by this system. If the answer is yes, then according to Gödel's second theorem it is not possible to prove with this cognitive system its own consistency, that is, a human cannot prove that her cognitive system is consistent. Given this, it follows that Lucas's argument is weakening to a great extent, because the argument is founded on a consistent system (machine).

(2) *Arguments based* *on* *the Chinese Room thought experiment.* Since I have referred to this thought experiment of Searle's (1980) in the previous chapters, I will now describe it in brevity. The fundamental idea is to show that even though a computer successfully passed the Turing test, it does not mean that the computer consciously understood what it was doing. All the computer does is act mechanistically, syntactically, on physical symbols without understanding their semantic meaning. Harnad (1990) call this the “symbol-grounding problem”, the question of how a symbol acquires its meaning. In the present case it is clear that while the Chinese understand the meaning of their symbols, the computer syntactic process alone cannot give Searle their meaning. Searle, who does not understand Chinese, takes it upon himself to act as a very successful computer in the Chinese language. He works on inputs in the Chinese language (Chinese characters) according to an English instruction book, and produces outputs in Chinese, which are judged by Chinese speakers to indicate a great understanding of the Chinese language. However, at the end of this thought experiment, Searle testifies that he does not understand a single word of Chinese. Hence, even the Chinese computer does not understand anything, even though it passed the Turing test with great success. This thought experiment provoked a very heated debate in the professional literature (e.g., Cole, 2023; Rakover, 1999). For example, Rakover (1999) developed a variation on the Chinese room that reinforces its conclusion. Accordingly, a computer named Golem imitates Searle in the Chinese room. Golem imitates the following sentence that Searle said repeatedly to himself in English, "I don't understand what these signs mean" (the Chinese characters). Now the following question arises here: does Golem understand English, which Searle understands (yes, no) and does Golem understands Chinese, which Searle does not understand (yes, no). An analysis of the four possibilities created here, leads to the clear conclusion that only the possibility: neither understanding English nor understanding Chinese, is the most likely possibility, a conclusion that disconfirms the hypothesis about a computer that thinks.

(3) *Arguments based* *on* *the viewpoints of Dreyfus*. Dreyfus' criticisms are directed more against the possibility of a computer simulating intelligent behavior, than the possibility of a computer developing CΨ. In other words, Dreyfus' criticisms are directed against the computer's explanations of cognitive states and processes rather than against explanations of how CΨ arises from processes in the brain. Despite this, there are two important reasons why his criticisms should be discussed here. First, Dreyfus' criticisms were directed against the conception of the mind as a device that manipulates symbols, processes information according to certain computational rules. In the 1960s this was an idea that was considered as the key to the door of understanding human cognition as a whole. Secondly, artificial consciousness (AC) is part of AI. In fact, Chrisley (2008) proposes that elaborations of AI may lead to AC.

Dreyfus' critique in the 1960s focused on the idea that human intelligence is not based totally on conscious mental manipulation of physical symbols. An important part of the cognitive processes is done unconsciously, and it would be very difficult to formulate these processes according to formal rules (e.g., Dreyfus, 1972; McCoruck, 2004). Dreyfus' critique was initially received negatively and was forgotten over time. In retrospect, it turns out that a significant part of his criticisms hit the mark. For example, the works of Tversky and Kahneman on judgment under conditions of uncertainty (which were conducted independently of Dreyfus' criticisms) showed that indeed humans act according to two different systems. System 1 is characterized as non-conscious, fast, intuitive and automatic. System 2 is characterized as conscious, slow, and thoughtful (e.g., Kahneman, 2011).

*New arguments against a conscious-robot*

Below I will briefly describe a number of arguments, thought experiments and story-like experiments (which describe in a fictional style imaginary experiments) that I have developed over time for the following reason. It seems to me that they are able to convince that indeed a conscious-robot has not yet been created and that the chances of such a device developing CΨ in the future are extremely slim.

(1) *The multi-functionality argument*. I use the term “multi-functionality” as a general concept that refers to the following phenomena: (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 is just another example of multi-functionality. One may conceive of a robot (or a computer) as one among other ways of imitating human behavior (motor, cognitive, mental). Now, given that many mechanical systems succeed in imitating human behavior, can one expect them all to develop CΨ and understanding? A positive answer seems like a huge exaggeration. Why? Essentially, because the attribution CΨ to robots and computers is based on analogy and similarity.

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 solve problems in logic and mathematics, answer general questions rationally, perform medical diagnoses, beat masters in chess and go, they can learn to speak, write poems, stories, music, paint, and so forth. So, do they have CΨ? My answer is no.

The fact that robots can imitate human behaviors so well does not necessarily imply that they can also develop CΨ and understanding. This is so because the hypothesis that a robot might develop CΨ 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 CΨ 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 CΨ 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 CΨ and understanding.

(2) *Anti-Complexity argument*. Here I am presenting one of the arguments that I developed against the idea of highly complex and sophisticated​​ robots with CΨ, the "Anti-Complexity" argument (e.g., McDermott, 2007). The basic idea is founded on two assumptions, which seem to me self-evident.

(a) *Human-like Artificial* CΨ *(A*CΨ) *may evolve in highly sophisticated and complex robots.*

It is clear to everyone that today calculators and even personal computers did not develop CΨ similar to that existing in humans. Therefore, the believers in robots with CΨ assume that the level of complexity and sophistication of the robots has not yet reached the threshold from which CΨ will appear in them. I emphasize here that to the best of my understanding the ACΨ that the believing researchers are talking about is a human-like CΨ, simply because I am unable to imagine another type of CΨ – non human-like CΨ (nor have I discover a description of another type of CΨ in the professional literature). In terms of Nagel (1974), I can say that I have no idea "what is like to be a machine" means. All I can think of is that it's not "what is like to be a human". Buttazzo (2001, p. 24) writes, “Many people now believe that artificial consciousness is possible and that, in the future, it will emerge in complex computing machines.”

(b) *Human CΨ is not related to the level of intelligence, knowledge, or the*

*complexity of the brain.*

The level of CΨ of the intelligent person is equal to that of the stupid person. Furthermore, a person's level of CΨ is not related to her level of understanding at different times. A person is aware that she understands as much as she is aware that she does not understand. Similarly, Koch (2019, p. 141) writes, “… intelligence and experience are distinct: being dumb or smart is different from being less or more conscious. … the neural correlates of consciousness with their center of gravity in the back of cortex are distinct from the correlates of intelligent behavior with their epicenter toward the front.” We discover CΨ in people who, for various reasons, different parts of their cerebral cortex have been severely damaged or have not developed (e.g., Merker, 2007; Nieder, 2021). Even animals, whose level of brain development is less than that of man, are in a state of CΨ.

*Conclusion*: If ACΨ is similar to human CΨ and if human CΨ is not related to the complexity and sophistication of the brain, then the hypothesis that robots will evolve human-like ACΨ when they reach the particular threshold of sophistication and complexity is false.

Of course, a believer in a robot’s ACΨ may suggest the following hypothesis: the above is true for humans but not of robots. Why? Because unlike humans, the development of CΨ in robots does depend on their degree of complexity. My response to this is as follows. Since no one knows what is the degree of complexity that gives rise to CΨ, the status of this hypothesis is questionable. The main reason is that this hypothesis does not allow for its empirical test (see argument 3 below).

(3) *The Robotic-Mom thought experiment.* Rakover (2023a) propose the “robotic-mom” thought experiment that argues that one who holds the hypothesis that a sophisticated robot may develop CΨ is in effect committed 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; Rakover1990). According to Popper (1959), it must be falsifiable; it must admit itself to the possibility of being refuted. As will be seen below, this requirement is not fulfilled by the “robotic-mom” story-like experiment.

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 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 story-like experiment is: what will the robotic-mom do when the nurse brings the robotic-baby to her? Two simple and opposing predictions, hypothetical responses, are possible.

*Prediction (A): hypothetical response (A)*. The robotic-mom will mimic the

human-mom exactly, i.e., she will reject the robotic-baby.

*Prediction (B): hypothetical response (B)*. The robotic-mom will *not* mimic the

human-mom’s behavior and will *not* reject the robotic-baby. *Why*? Because *if* she has developed CΨ, *then* she will understand that the robotic-babyis her own kind and is her real baby. Furthermore, if the human-baby were brought to her now, she would reject him. 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.

Although there is no empirical evidence concerning these two hypothetical predictions, the scientific status of the two hypothetical predictions can be evaluated. It is clear that *hypothetical prediction (A)* meets the accepted requirement of a scientific hypothesis, refutability. Accordingly, if the robotic-mom does *not* reject the robotic-baby, but accepts it warmly, *prediction (A)* will be refuted.

This situation does not hold for *prediction (B): hypothetical response (B)*. According to this prediction, if the robotic-mom develops CΨ and understanding, she will not elicit response (A), but will elicit response (B), i.e., she will accept the robotic-baby. However, what happens if the robotic-mom rejects the robotic-baby? Is *prediction (B)* refuted? The answer is no! Why? Because it is possible to propose that the robotic-mom has not yet reached the critical stage of advancement and sophistication that creates CΨ and understanding. The problem with this argument, which on its face seems attractive and rational, is that it makes this prediction an ad hoc hypothesis. Why? Because it can be invoked whenever this prediction has been refuted. One may use this ad hoc argument as an excuse whenever the robotic-mom, no matter how advanced and sophisticated it is, does not accept the robotic-baby. Therefore, *prediction (B): hypothetical response (B)* is immune to empirical testing, to the possibility of refutation. It can be seen as a mere faith and not as a scientific hypothesis.

(4) *CΨ-Counter*. Rakover (2023b) proposed the “CΨ-Counter” story-like experiment to emphasize the possibility that the scientific methods, which were developed in the natural sciences and imported to psychology, may not be adequate for research in CΨ. A group of scientists develops a new theory of CΨ and a device called the “CΨ-Counter”. The device can measure any kind of CΨ in objective measurement units (OMUs), and it can also measure in “OMUs” any physical or chemical property. This device was applied in two cases. First, when Mrs. Smith from New York, a lover of Renaissance art, saw the Mona Lisa and her level of “art excitement” as measured by the CΨ-Counter was +187 OMU. Secondly, in an art survey, it was discovered that an environmental sculpture installed in Paris, made of objects that had been discarded and retrieved from the municipal garbage dump, irradiated exactly +187 OMU. Therefore, it was hypothesized that Mrs. Smith’s impression of the Mona Lisa would equal her impression of the environmental sculpture, +187 OMU precisely. However, when Mrs. Smith was flown to Paris and saw the sculpture, the CΨ-Counter recorded -273 OMU. She detested the sculpture immensely. Thus, the present hypothesis was disconfirmed. Therefore, one may propose that the theory and the CΨ-Counter depended on a methodology developed for investigating the external, natural world and is not appropriate for research of the inner world – human CΨ.

(5) *Live-creatures’ correlation*. The present hypothesis, called “Live-creatures’ correlation”, is based on an empirical generalization, which is supported by the following observations (see chapter 6). Only living beings that are made of organic matter and that have a certain system of brain and nerves (as in humans and animals) have CΨ. The observations propose that only living beings such as humans, monkeys, dogs, cats, dolphins, etc., are endowed with CΨ, while inanimate beings such as plants, soil, stones, buildings, cars, toasters, etc. are not endowed with CΨ (and as I stated in the previous chapters, I do not believe in panpsychism). Buttazzo (2001, p. 27) writes, “Although most people can accept the existence of holistic features [CΨ is a holistic property], others still believe that consciousness cannot emerge from a silicon substratum because it is an intrinsic property of biological materials such as neural cells.” Additionally, Blackmore (2013) discusses the possibility that CΨ is limited only to biological creatures, but undermines the possibility that there is a functional difference between, for example, the neurons in the brain and the artificial means that imitate them.

If indeed the live-creatures' correlation holds, then it can be said that robots and computers, which are made of material that is different from animals’ stuff, do not have CΨ. The question is of course whether this correlation holds empirically. If yes, then the conviction that robots and computers will not develop CΨ is getting stronger. If not, since it is discovered, for example, that even stones have tiny elements of CΨ or that plants have a certain human-like-plant-CΨ, because they absorb environmental stimulation, then the support of this correlation, which suggest that the existence of a conscious robot is impossible, will be greatly weakened.

(6) *The Genetic-Attribution hypothesis*. In light of the above, it can be said that the decision whether a robot is endowed with CΨ is not based on conclusive proof, but on a certain degree of conviction. In the current section, I will describe a hypothesis that suggests the following. It is relatively easy to attribute CΨ to humans, it is difficult to attribute it to animals, and it is very hard to attribute it to robots or computers. The hypothesis, which I call the "genetic-attribution” hypothesis, is founded on the following idea. Because I am the child of my parents, a product of their qualities, then it is logical to assume that if I am endowed with CΨ so are my parents – they transfer it to me. Thus, it is easy to generalize this conclusion, to my family (brothers, sisters, uncles and aunts and their children), to friends, members of the tribe, the nation, to human beings. It is more difficult to generalize CΨ to animals, unless one believes in the Darwinist theory that proposes a common origin for all live creatures (so all animals have CΨ). Given this rationale, it is clear that it is very difficult to attribute CΨ to robots and computers. Is the genetic-attribution hypothesis supported? Let us examine it in the light of the following story-like experiment, which I will call "Einstein Mind". It demonstrates the problematic nature of attributing CΨ to humans and robots.

Suppose that a “certain creature” that resembles Einstein like two drops of water behaves exactly like Einstein. Who is that ‘certain creature’: a robot or an actor? To whom would one attribute Einstein's mind, to (A) a highly sophisticated robot or to (B) an extremely gifted theater actor. I guess that the answer is (A) because it is clear to everyone that (B) would never have been able to come up with the special and general relativity theories.

In conclusion, I hope I have managed to discuss here as fairly as possible the main arguments for or against the possibility of a machine-CΨ, despite my conviction that machines have not yet developed CΨ and the chances that they will develop CΨ are extremely slim. I believe that the reader himself has to decide in what he/she is convinced of (including the situation in which he/she decides not to be convinced). Anyhow, it seems to me that the following story-like experiment may add a special color, touch, to the current chapter. Here it is.

George Pygmalion is a robotics engineer who has spent years building a robot named Gigi Galatea. He especially developed in her the trait of empathy, compassion and sincere participation in human suffering. One day he entered his laboratory early in the morning and immediately started working. All of a sudden, he screamed in pain. "What a terrible blow I received in the hand, oh, how much it hurts!" He opened and closed his fist to check if any bones were broken and continued to scream, "Oh, it hurts like hell."

"Don't worry, my dear," Gigi said in a warm, compassionate and empathetic voice, "Don't worry, my dear, you know, I too got a strong blow on my hand and the pain passed very quickly. It will pass very quickly for you too, don't worry."

"What are babbling, Gigi," said George Pygmalion in astonishment mixed with anger, "what are you talking about, I haven't put your hands on your body yet."

**References**

Avramides, A. (2023). Other Minds. In E. N. Zalta & U. Nodelman (eds.), *The*

*Stanford Encyclopedia of Philosophy*, URL = <https://plato.stanford.edu/archives/win2023/entries/other-minds/>.

Blackbum, S. (Ed., 2016). [Ship of Theseus](https://archive.org/details/oxforddictionary0000blac_n2u8). *The Oxford dictionary of*

*philosophy (3ed)*. Oxford: Oxford University

Blackmore, S. (2013). *Consciousness: An introduction*. London; New York:

Routledge.

Buttazzo, G, (2001). Artificial consciousness: Utopia or real possibility?

*Computer,* 0018-9162/01/$10.00 © 2001 IEEE.

Chalmers, D. (1996). *The conscious mind: In search of a fundamental*

*theory*. Oxford: [Oxford University Press](https://www.wikiwand.com/he/Oxford_University_Press).

Chella, A. and Manzotti, R. (2011). Artificial consciousness. In V.

Cutsuridis et al. (eds.), *Perception-action cycle: Models, architectures, and hardware.* Springer Series in Cognitive and Neural System 1, Springer Science + Business Media.

Chrisley, R. (2008). Philosophical foundations of artificial consciousness.

*Artificial Intelligence in Medicine,* 44, 119-137.

Cole, D. (2023). The Chinese Room argument. In E. N. Zalta & U.

Nodelman (eds.), *The Stanford Encyclopedia of Philosophy*. URL = <https://plato.stanford.edu/archives/sum2023/entries/chinese-room/>

Dreyfus, H. (1972). *What computers can’t do: A critique of artificial reason.*

N.Y.: Harper & Row.

Elsey, J. and Moss, D. (2023). Public opinion of AI policy and risk. May 12,

2023, Rethink Priorities.

Faverio, M. and Tyson, A. (2023). What the data says about Americans views

of artificial intelligence. Short reads/November 21, 2023, Pew Research Center.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear*

*Phenomena,* 42, 335-346.

Kahneman, D. (2011). *Thinking, fast and slow.* New York, NY: Farrar,

Straus and Giroux.

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

2793

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

Koch, C. (2019). *The feeling of life itself: Why consciousness is widespread but*

*can’t be computed.* Cambridge, Mass.: The MIT Press.

Kurzweil, R. (1999) The Age of Spiritual Machines, New York and London,

Texere.

Lucas, J. R. (1961). Minds, machines, and Gӧdel. *Philosophy,* 36, 112-127.

McCorduck, P. (2004*). Machines who think (2ed)*. MA: A. K. Peters.

McDermott, D. (2007). Artificial intelligence and consciousness. In P. D.

Zelazlo, M. Moscovitch and E, Thompson (eds.), *The Cambridge Handbook of Consciousness.* (Pp. 117-150). Cambridge: Cambridge University Press.

Megill, J. (2024). The Lucas-Penrose argument about Gӧdel’s theorem.

*Internet Encyclopedia of Philosophy.*

Merker, B. (2007). Consciousness without cerebral cortex: A challenge for

neuroscience and medicine. *Behavioral and Brain Sciences,* 30, 63-134.

Milkowski, M. (2024). The Computational Theory of Mind. *The Internet*

*Encyclopedia of Philosophy*. ISSN 2161-0002, <https://iep.utm.edu/>

Müller, V. C. (2024). Philosophy of AI: Structured overview. In N. Smuha

(ed.), *Cambridge Handbook on the Law, Ethics and Policy of Artificial Intelligence*. Cambridge: Cambridge University Press.

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

435–450.

Nieder, A. (2021). Consciousness without cortex. *Current Opinion in*

*Neurobiology*, 71, 69-76.

Nola, R. & Sankey, H. (2007). Theories of scientific method: An introduction.

Stocksfield: Acumen Publishing.

Penrose, R. (1989). *Emperor’s new mind.* Oxford: Oxxford University press.

Penrose, R. (1994). Shadows of the *mind.* Oxford: Oxxford University press.

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

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. (2023a). Is a Conscious Robot a Scientific Hypothesis or Just a

Faith? *The Journal of Mind and Behavior*, 44, 147–156.

Rakover, S. S. (2023b). What will happen if science will develop a theory of

consciousness? Negative Ramifications. *Journal of Conscious Evolution*, 17, 4-10-2023.

Rescorla, M. (2020). The Computational Theory of Mind. In E. N. Zalta (ed.),

*The Stanford Encyclopedia of Philosophy*, URL = <https://plato.stanford.edu/archives/fall2020/entries/computational-mind/>

[Russell, S. J.](https://en.wikipedia.org/wiki/Stuart_J._Russell) and [Norvig, P.](https://en.wikipedia.org/wiki/Peter_Norvig) (2003), *Artificial intelligence: A modern approach*

*(2ed)*. Upper Saddle River, New Jersey: Prentice Hall

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences,*

3, 417-457.

Selmer, B. and Govindarajulu, N. S. (2024). Artificial intelligence. In E. N.

Zalta & U. Nodelman (eds.), *The Stanford Encyclopedia of Philosophy. URL= <https://plato.stanford.edu/archives/fall2024/entries/artificial-intelligence/>*

Vallance Chris, (13 June 2022, BBC). Google engineer says Lamda AI system

may have its own feelings