Do Robots Have Consciousness?

Agastya Brahmbhatt
Mr. Jack Bowen
Philosophy

Have you ever considered the possibility that robots might possess consciousness? While biological evolution has undeniably produced conscious, sentient beings from mere atoms and molecules, the question of whether a system of electricity and hardware could achieve the same remains unresolved. From a scientific standpoint, human free will is debatable, leading to the provocative idea that we are, in essence, "biological robots," whose actions are dictated by genetic programming and environmental factors, much like a robot's functions are determined by its CPU. But why should this question of robotic consciousness matter? A deeper understanding of robot consciousness has significant implications for how we prepare for the progression of artificial intelligence, and possibly, for the prospect of a future dominated by robots. Regardless of whether robots truly achieve consciousness, exploring this concept can shed light on the nature of consciousness itself and its impact on moral decision-making, as well as shape our broader understanding of our place in the universe.

The metaphysical viewpoint from Plato and Aristotle held that one died because their soul left their body, so the soul is responsible for consciousness and the mind. The primary counterargument is that robots do not have consciousness, as consciousness is merely due to biological phenomena, and devoid of a soul. Electricity and python programs are non-biological, and so therefore cannot be sentient or conscious. Thus, one can argue that every robot is mindless. In contrast, one can interpret that robots have consciousness because of Descartes's viewpoint on the soul's translation to a mind, science's input on what consciousness is, and the exploration of ontology versus epistemology by contrasting humans and robots. The reality or definition of a robot or a prosthetic human or a cyborg will also change along with our understanding of neuroscience trying to explain what consciousness is. Synthetic humans, cyborgs, and minds in petri-dishes are other applications of robotic consciousness. The debate as to whether robots have consciousness poses many existential questions, and whether A.I. consciousness exists.

The metaphysical viewpoint from Plato and Aristotle argued that one died because their soul left their body, so the soul is responsible for consciousness and the mind. From ancient times, humans have always wondered whether there is something special about their consciousnesses, existence, and sense of being. The development of the human mind is such that it is impossible to imagine no consciousness, or being "passed" or "knocked out," as it is impossible to feel how it feels to feel nothing, or to sense no sensation, such as nothingness after death and before birth. Since humans have not been able to imagine such a feeling, they have always wanted to fill the void with afterlife, god, soul, and theological ideas, no matter how small and remote an island where a civilization developed. The earliest written accounts of such ideas are well-documented in the works of Plato, Aristotle, and other Greek philosophers. Plato believed that animals, plants, etc. all had souls. According to Poole,

> There is more nuance in how Plato sees the soul [...] it is not all sublime. Plato thinks it has three parts: the base part, which plants also have, is appetitive and concerned with

> desire, pleasure, and pain. The middle part of the soul, which animals also have, is the spirited part of the soul which seems to act as a sort of charioteer [...] sorting out the appetitive part on behalf of the highest part of the soul. The highest part of the soul is peculiarly human and is the rational or intellectual part of the soul (Poole, 2024).

Plato believes that the soul is divided amongst three subgroups, the base part, the middle part, and the highest part. In Plato's idea of a soul, different forms of consciousness depend on a hierarchy, spreading from plants to animals to humans. Biologically, these three have the common feature of life, while other inanimate objects such as books and tables do not. It is then plausible to earlier philosophers that anything that has life should thus have a soul. However, despite the fact that all life forms have souls, some life forms have "higher" soul parts than others. For example, the anthropocentric ability to rationalize is considered higher than animal-like, impulsive emotional-based behavior. In contrast to Plato, Aristotle believed in the anthropocentric separation of mind from pure animal-like emotion. He also believed that souls were essential to our unique qualities. According to Poole,

> Aristotle devoted a whole treatise to the soul. As for Plato, Aristotle believed that plants and animals had souls, although humans were the only entities that had a rational soul, or 'mind', which is both immaterial and eternal [...] while he disagreed with Plato about the universality of forms, he argued that in an individual, our soul is to our body as form is to matter. In his thinking, our soul is proprietary to us, not a manifestation of something universally generic (Poole, 2024).

The Aristotelian belief of the soul is that humans have a distinct rational mind, while animals do not. (It is interesting to note that some animals do have rational qualities, yet not at the level of human logic.) One way Aristotle diverges from Plato is the belief that the soul is linked to one's uniqueness and individuality. However, an interesting question arises when you consider a human being who is brain-damaged, or with poor rationale. Would such a human not have a soul? Why is it that rationality and the capacity to have a mind can make humans special? Robots can now 'rationalize' in the sense that they can do complex rational scientific and mathematical problems within milliseconds, far beyond a human's rational ability. In fact, robots can also write philosophical papers. Recently, due to ChatGPT's passing of the Turing Test, it would seem that there is no notable difference between human rationality and robot rationale programming. In this case, Aristotle would have been unable to claim that robots cannot have minds. Descartes believes that the soul is not something that affects the body, but the other way around. According to Poole,

> [Descartes] agrees that it is true that the soul departs on death, but because the body has died, not because the soul's departure is making it dead. It is this last development in thought that turns the future discourse from souls to minds and

> consciousness, and arguably away from the theologians and towards the philosophers [and scientists] (Poole, 2024).

Descartes believes that assigning the soul the job of all bodily endeavors is too implausible. It seems that the soul is at the mercy of the body. Such thought led to the Aristotelian mind, and then to the modern understanding of scientific consciousness. It seems that our consciousness is at the mercy of our biological programming, as a robot is at the mercy of their program. However, at what point does the robot become conscious? We agree that basic biological forms like cells aren't conscious, but they are still biological, so biological doesn't necessarily mean conscious. Similarly, just because robots aren't biological doesn't mean they can't be conscious.

Robots are mindless because robots do not have consciousness, as consciousness is merely an illusion of biological phenomena. Electricity and python programs are non-biological, and so therefore cannot be sentient or conscious. Philip Ball argues that every robot is nothing but electrons moving, unlike our basic biology. It is simply too electronic to be conscious. According to Poole, "Let's be clear: every robot and computer ever built is mindless [...] I might be merely repeating the bias against animal minds that I have" (Ball, 2022). It certainly seems believable. How can a steel hunk of metal be conscious? Robots are inanimate objects, and so therefore by definition they should be devoid of a consciousness, mind, or soul. However, when thinking about this deeper, it seems that the argument for why robots aren't conscious also pertain to ourselves as well. If we agree with John Searle that the mind is not a physical entity, then it seems odd to definitively state that robots can't have a consciousness, as they are just hunks of metal. A mind could still exist as a product of the metallic, silicon product. In silicon valley, data centers containing metal and silicon parts can pass the Turing Test. The data center's "mind" is also not a physical entity, but the future seems increasingly plausible towards an even more sophisticated version of artificial intelligence, that may pose certain moral and ethical questions and dilemmas. Organizations and their employees have a moral responsibility to conduct their business in ethical and legal fashion. Any violation does not imply that the organization itself should go to trial and go to jail. Similarly, pets may acquire violent habits and may be put to death. However, there is no jail for pets. Similarly, Ball's argument is that when a machine such as a robot starts malfunctioning, it should be discarded. Today, an automated driving car accident may imply either the fault of the car company, or the driver. It would not be the car itself. However, can the robots in the future be advanced enough to violate a law or misbehave? Would such robots be morally accountable? According to Ball, "Dennett proposes that giving a machine consciousness would create a burden we could do without, because we ought then to afford it moral rights" (Ball, 2022). Dennett's analysis of giving robots moral rights is quite similar to giving moral rights to people without free will, except that robots can be manually programmed and fixed to not commit such morally atrocious behavior. If a serial killer used an excuse in trial that he was devoid of free will, it would still be customary to lock him up due to utilitarianism, as the freedom of a serial killer is far outweighed by his victim's deaths. However, in the case of a robot, if a robot misbehaves, no trial would be necessary as

programmers could easily fix the bug that led to the robot's initial moral downfall. This arises because the robots can be fixed, while we have still not found ways to rehabilitate psychopaths. Just because we don't want to put robots on trial, however, doesn't mean that we should assume that they will never be conscious. A.I.'s are already affecting human behavior with polarized social network contents, automated driving and ChatGPT. According to Ball,"Machine behavior represents [...] striking examples of the dictum [...] 'we shape our technologies, and then our technologies shape us'" (Ball, 2022). More A.I. tasks, including medical support, imply that A.I. is and will be an even more integral part of human life. Therefore, A.I. is becoming more human and should therefore have consciousness to benefit humanity.

In contrast, one can interpret that robots have consciousness if one believes in Descartes's opinion on the soul, the scientific viewpoint on consciousness, and the contrast between ontology versus epistemology (discussing the remaining differences between humans and machines). One article that explains the modern theory of the soul while complying with basic scientific knowledge is Thomas Nagel's famous article on the epistemological block regarding our knowledge of bats. What we understand as a brain and sensory organs are one unit that receives a stimulus, acquires a memory, processes the stimulus, and reacts to it. As we have pointed out elsewhere, if a robot was a black box with similar behavior, it would not be possible to tell whether this processing and reaction to the stimulus came from a robot or a human. Similarly, when and how a robot may sense the stimulus and react to it seems different compared to humans. If a robot's behaviors were unpredictable in its reaction and complex enough, it is difficult to justify why it is not reacting as an individual. We do not differentiate one bat from the other, even though we may understand different individual characteristics of our pets differently. However, we still believe that bats have consciousness. So why should robots be any different? Humans did not believe that robots could hold as much knowledge as a human. Robotics, as a field that is maturing, has now passed the Turing Test. Through neural networks, Robot minds are becoming more identical to human minds in their ability to possess knowledge. What now remains is the individual experience unique to humans. According to Poole,

> Ontology is what we know and epistemology is how we know it. This is important in our discussion about artificial intelligence, not only because we are trying to understand [...] persons and consciousness 'ontology, but we are trying to clarify [...] A.I. can theoretically replicate our own (Poole, 2024).

Since the artificial intelligence that has passed the Turing test sits in a giant data center, we can easily claim that it does not have any experience that a human experiences in a society. However, that should not stop future robots or devices from acquiring their own artificial neural networks that will learn and become individuals based on their own unique experiences. One may end up having a robotic pet that learns to recognize only its owner and act as a therapist and have a unique personality. What still differentiates such machines from real pets is that they do not have a nervous system and do not experience pain or other neurological phenomena. However, why

should a future robot pet not truly experience a feeling of pleasure to see its owner at the end of the day?

Poole wants to differentiate between humans and robots by bringing out the key differences in human behavior, which pure logic doesn't support. According to Poole, "Junk code in our own programming, [...] instead of being dismissed, [...] should be articulated, nurtured, and protected: emotions, mistakes, uncertainty, storytelling, free will, sixth sense, and meaning" (Poole, 2024). It is true that we have emotions, we make mistakes, and many times we are uncertain about our decisions. We do use the term "human error" to imply the human element in occasional mistakes. However, the claim that humans behave with certain properties in their behavior should not imply that only beings with those properties are capable of having a soul. A dog or a cat does not have storytelling or a perception of meaning in a human-like sense. However, we do not say that these animals are not conscious. Poole wants to celebrate the human weaknesses and limitations and she should be allowed to do so. However, she wants to imply that we have created logic and machinery capable of overcoming those weaknesses as a reaction to the knowledge of human limitations. She proposes that we should convert the future robots to behave in a more human-like fashion. According to Poole, "I think this is the pivot we need to make in A.I. The first phase has been about stripping out all the junk code [...] now we need to think about programming all of that junk code back in [...] because actually, that is us at our best" (Poole, 2024). A general observation about robots across the philosophical works that I have come across is that robots either compete, or they complement the humans. What Poole is proposing is neither competition, nor the complementing of human limitations, but imitation. One classic example is spell-check error correction used by early A.I. in a search engine. It is difficult to imagine that humans would be thrilled if they typed a correct spelling and the search engine reacted as if it was an incorrect spelling and gave it wrong results occasionally. There are ample examples of industrial products which are not functioning on a predictable basis and failing to succeed because humans do not want to celebrate mistakes or whimsical behavior in a robot or machine. Poole's proposal may actually lead to a disturbing scenario of human-like robots with similar tendencies roaming among human populations. At that point, one may wonder if there is any need for such machines.

The definition of a robot or a prosthetic human or a cyborg will also change along with our understanding of neuroscience trying to explain what consciousness is. Humans have been fascinated by the idea of human-like robots for a long time. However, the reality of artificial intelligence is far more complex and intertwined with humans than what science fiction writers or philosophers in the early years of computer evolution imagined. For example, during Andrew Vate's era, computers were clunky, large, expensive, and only seen in a laboratory setting. According to Vate,"to call the computer "theoretically human" is no better than to call it "metaphorically human." (Vate, 1971). The idea that a computer can fit in one's pocket and listen and learn from its owner was unimaginable. Today, not only can a computer do all of those activities, it is connected to an artificial brain. In some sense, it is both learning from human experience and collecting all the data from the entire humanity in one place and processing it.

There is no reason to believe that a future artificial "brain" will not be mobile or have many mobile sensors and it could acquire more sensory dimensions besides visual images and sound files.

Synthetic humans, cyborgs, and minds in petri-dishes are other applications of robotic consciousness. Scientists have already developed biological neuronal brains in petri-dishes with the help of STEM cells and these "brains" are already capable of recognizing basic patterns and behaving differently in the presence of unknown stimuli. If such brains become a part of a prosthetic human, how would we be able to tell which portion of that brain is robot and which is human? How do we convince ourselves that these "brains" in the petri-dish are not secreting serotonin or dopamine, and are not feeling pain or consciousness? According to Putnam, "there is no correct answer to the question: are robots conscious?" (Putnam, 1964). Even though Putnam's question and conclusion were based on a hypothetical robot, the prosthetic human and brains in the petri-dish may really put a question mark to what is a living organism. If a petri-dish brain is sentient, would it be a crime to inflict pain or kill it? With quantum computers and analog electronics, the idea of artificial intelligence will not remain as an abstract computer program as Searle viewed it. It may even get attached to a human being as a body organ or his or her brain. We have no choice but to ask the way Thomas Nagel asked, "What is it to feel like a petri-dish brain?"

To sum up, the theological viewpoint from Plato and Aristotle argued that one died because their soul left their body, so the soul is responsible for consciousness and the mind. The primary counterargument is that robots do not have consciousness, as consciousness is merely an illusion of biological phenomena. Electricity and python programs are non-biological, and so therefore cannot be sentient or conscious. Thus, one can argue that every robot is mindless. In contrast, one can interpret that robots have consciousness because of Descartes's discussion of how the soul translates to minds and consciousness, science's input on the consciousness discussion, and the discussion of ontology versus epistemology (discussing the remaining differences between humans and machines). The reality or definition of a robot or a prosthetic human or a cyborg will also change along with our understanding of neuroscience trying to explain what consciousness is. Synthetic humans, cyborgs, and minds in petri-dishes are other applications of robotic consciousness. Overall, the mystery as to what lies at the root of all human consciousness has still not been answered. Whether we can apply our estimates of consciousness' root accurately in the case of answering whether robots have consciousness is a separate question. I believe that until we know what exactly is at the root of all human consciousness (some form of what we define as a soul), we cannot answer the question as to whether robots have consciousness, because we still have no understanding of the source of consciousness. Thus, the best we can do is to compare ourselves to the robots. Without free will, it seems impossible to make any more distinctions between us and robots than the fact that we are biological, and we know that we are sentient and experience emotion. However, emotions are just chemical reactions, and feelings aren't tangible. One might argue that the metaphysical is purely pseudoscience that doesn't exist, so it seems that the root of all intangible sentience is

material cause. Thus, if robots have material causes powering their actions and circuits, then I too believe that there is a good likelihood that robots have a similar illusion of consciousness that we do, at least the complex robots. In the end, is there really a definitive answer? It's hard to tell.

Bibliography

Ball, Philip. *The Book of Minds: How to Understand Ourselves and Other Beings, from Animals to AI to Aliens*. Chicago: University of Chicago Press, 2022.

The author, Philip Ball, is a philosophical writer at the University of Chicago press. His works range from *Bright Earth* to *the Elements.* He won the 2005 Aventis Prize for Science books. He was a chemist at Oxford and an editor for *Nature.* The source claims that in this universe, there is an existence of a space of possible minds. Such include the minds of animals to the minds of A.I. If we assign personhood to one mind, we must do so to the other. This source answers whether robots have souls or not by explaining that we are not quite different from biological robots. If we assign personhood to ourselves, then why not an artificial one? The evidence that the source uses is examples ranging from animals to artificial intelligence. Anything that is sentient or has the illusion of a consciousness/free will must also deserve personhood, because there is no way to prove that we, as people, are sentient and aren't bots with the illusion of sentience. This source isn't biased because it takes into account the perspective of artificial machines. It is quite non anthropocentric, and trying to take into account all of the perspectives ranging from animals to microorganisms, to robots. The source addresses the counterargument that robots don't have souls by posing questions as to how we are any different from robots. The logic of the argument is sound, and there are a minimal number of gaps to be addressed.

Gamez, David. "Machine Consciousness." In *Human and Machine Consciousness*, 1st ed., 135–48. Open Book Publishers, 2018. http://www.jstor.org/stable/j.ctv8j3zv.14.

The author, David Gamez, is a professor at the University of Essex. His many works range from philosophy books titled *What We Can Never Know* to editorials by Open Book publishers. The source argues that robots can only be conscious and can have the capacity of thought only if they have the same power as brains. This source answers the question of whether robots have consciousness or not because it explains how robots have consciousness because some machine-like systems can display conscious behavior and can have bubbles of experience, and one can build CC sets to model consciousness. The evidence the source uses to support its argument is that the mind is beyond immaterial processes. The other evidence used is that computer scientists have been using multiple methodologies to harness robot consciousness over many years. The source may be a little bit biased, as there is no evidence that such methodologies have been successful yet. There isn't a large variety of evidence. The source rarely addresses the counterargument. The logic of the argument is sound, but the gaps to be addressed are that minds may be immaterial, but how sure can you be that they don't result from biological processes? If a robot is devoid of such biology, then how can it have a mind?

Nahmias, Eddy, Corey Hill Allen, and Bradley Loveall. "When Do Robots Have Free Will?: Exploring the Relationships between (Attributions of) Consciousness and Free Will." In *Free Will, Causality, and Neuroscience*, edited by Bernard Feltz, Marcus Missal, and Andrew Sims, 338:57–80. Brill, 2020. http://www.jstor.org/stable/10.1163/j.ctvrxk31x.8.

The author Eddy Nahmias is a professor of philosophy at Georgia State University. He has written books ranging from *Moral Psychology* to the *Natural Method*. The primary thesis answers the question of whether or not humanoid robots are morally accountable. The thesis is that assuming that they don't have free will, then how can they be morally accountable? If a robot consciousness has no free will, then it can't be morally accountable. However, it would be necessary to lock them up if they did any wrong, due to utilitarianism. The source answers the question of whether or not robots have consciousness by answering points of whether robot consciousness implies free will, personhood, and moral accountability. The evidence the source uses to support its argument is through observation of the moral accountability given to humans. It also assesses robotic moral accountability, but there is limited evidence of whether or not humanoid robots could ever be morally responsible. However, the evidence it uses depends on whether or not humans could have free will in a robot's body. In other words, it depends how they are designed. The source isn't biased or leaves out external perspectives because it hasn't left much evidence unturned. There is plenty of evidence and a variety of evidence. The source addresses the counterargument by analyzing whether or not a lack of free will can ever imply moral responsibility. The logic of the argument is sound and there are no gaps to be addressed.

Poole, Eve. *Robot Souls: Programming in Humanity*. Boca Raton, FL: CRC Press, 2024.

Eve Poole is a British writer and CEO of the Carnegie Trust and Edinburgh Royal Society. She is known for her works ranging from *Buying God* to *Capitalism's Toxic Assumptions.* The source's thesis is that Robots have souls because it poses the question of why we should be paranoid about robot takeover, and why we should work out who we want to be, and how we want to characterize our relationship with A.I. The source answers the question that robots have souls by posing questions such as what is consciousness, and how can we relate to A.I.? The evidence the source uses stems from theories from artificial intelligence to the mind's eye. The source isn't biased. There is enough evidence and a variety of evidence to support it. The source does address the counterargument by posing questions on consciousness. The logic of the argument is sound. There aren't many gaps left to be addressed.

Vate, Dwight van de. "The Problem of Robot Consciousness." *Philosophy and Phenomenological Research* 32, no. 2 (1971): 149–65. https://doi.org/10.2307/2105945.

The author was the president of the Southern Society for Philosophy and Psychology. His works include *Romantic Love* to *The Goffman Lectures.* The thesis is that robot consciousness poses many issues because if a conscious flesh and bone entity is a social achievement, then would the conscious machine be a possible social achievement? The source answers the question of whether robots have consciousness or not by posing many questions and identifying issues as to whether or not the creation of a robot consciousness would be considered an achievement, or a disaster, potentially leading to an apocalypse. The evidence that the author uses ranges from theories of artificial intelligence to whether or not artificial intelligence enhancement is exceeding the level at which we are able to fully grasp and control its complexity. The source may be a little biased because it doesn't talk of the potential benefits of robot consciousness, and that it could lead to social achievements, not social drawbacks. There is enough evidence to support it, but not a variety of evidence addressing the counterargument. However, the source does address the counterargument a bit, that robot consciousness may pose a social achievement. The logic of the argument is sound, and there are no gaps left to be addressed.

Putnam, Hilary. "Robots: Machines or Artificially Created Life?" *The Journal of Philosophy* 61, no. 21 (1964): 668–91. https://doi.org/10.2307/2023045.

Hilary Putnam was a philosopher and computer scientist best known for his studies on the philosophy of mind, math, and science. A few of his famous books include *Mind, Language, and Reality* as well as *Realism With a Human Face.* The source's thesis is that we should be concerned about whether robots have souls or not, as it brings forth Wittgensetin's private language argument. The source answers the question by addressing further questions towards answering the mind-body problem, as well as the problem of minds of machines as well as logical behaviorism. The source uses Wittgenstein's theories as well as other theories surrounding the mind-body problem to answer the mind-machine question of whether robots have souls. This source may potentially be biased, as there is a lack of variety of evidence except for Wittgenstein. There is enough evidence supporting his conclusions, however. The source does address the counterargument in multiple ways, such as posing questions related to logical behaviorism. The logic of the argument is sound. There are minimal gaps left to be addressed.

Secondary Sources:

Chella, A., Cangelosi, A., Metta, G., & Bringsjord, S. (2019). Editorial: Consciousness in Humanoid Robots. Frontiers in robotics and AI, 6, 17. https://doi.org/10.3389/frobt.2019.00017

Antonio Chella is a researcher and philosopher from the University of Palermo and is best known for his books *Computational Approaches to Conscious Artificial Intelligence* as well as *Artificial Consciousness.* The source's thesis is that understanding robot consciousness is vital to understanding biological consciousness. The source answers the question of whether robots have these by posing biological applications of understanding robotic consciousness, and how these relate to our understanding of ourselves. The author uses evidence related to the brain as well as neural networking systems. It is not biased, as there is plenty of evidence on both sides of the argument mentioned. There is enough evidence and a variety of evidence to support the counter argument as well. The logic of the argument is sound. There are no gaps left to be addressed.

Kanai, Ryota. "We Need Conscious Robots." *Nautilus*. Last modified April 20, 2017. https://nautil.us/we-need-conscious-robots-236579/.

Ryota Kanai is a neuroscientist and philosopher. He is known for being the CEO of Arya, inc, which builds intelligent robots. He was also a former associate professor at the University of Sussex. The source's thesis is that conscious robots are necessary for understanding the role the consciousness plays in understanding more about the universe as well as ourselves. The source answers the question by figuring out how to program consciousness, as well as inventing before discovering all of the laws first. Rather than question and go in philosophical circles, inventing the conscious robot seems more important. The evidence that the source uses is attempting to justify building a conscious robot rather than not doing so. The evidence primarily relies on the experience of the industry of robotics as well as understanding where artificial intelligence may lead to in the future. The source may be a bit biased, because it doesn't mention the negative repercussions of robot consciousness as well as addressing the counterargument, which is that philosophical analyzation should take precedence over scientific invention. There may not be enough evidence/variety of evidence to support his claims. The counterargument is not addressed in any way. The logic of the argument is sound, but lacking in evidence and answering the ought vs. can question. There are no other gaps left to be addressed.