Jeremy Kalfus

Mr. NAME

Philosophy of Mind

15 November 2024

# On the Assessment of AI Consciousness

Throughout the history of philosophy of mind, consciousness has often been ascribed to humans only, though in recent times it has increasingly been expanded to animals of higher order. In contemporary thought, consciousness is seen as a phenomenal property that can be explained as the feeling of "what it is like" to be something, or in some state (Nagel 436). This concept is known as qualia. The specific range of what species of animal life is considered conscious has primarily been increased due to scientific development, where novel research has found shared physical and behavioral qualities between humans and the animals in question. Such a viewpoint implies that consciousness is chiefly biological—a statement that went without question for much of history—after all, one would likely be ridiculed for saying something such as a grain of salt is conscious. But from the start of the information era, it has become increasingly foreseeable that complex computational programs could be created with complexity that rivals the human mind. These programs are termed Artificial Intelligence (AI). Philosophy of mind generally classifies theoretical AI agents into two classes, Strong AI and Weak AI. Weak AI is artificial intelligence that is not conscious and performs "narrow" tasks that "would be considered intelligent if carried out by human operators" (Bory et al. 2). ChatGPT, AlphaFold, and self-driving cars are all considered examples of Weak AI. Unlike Weak AI, Strong AI is a hypothetical artificially intelligent agent that does not just emulate or claim to be conscious but

actually is. While current models of consciousness are adept in giving possible explanations for the phenomenon, they are incapable of assessing whether or not AI can be conscious due to limitations in their conceptual frameworks and the fact that they are built around the human mind.

Functionalism is a framework for understanding consciousness that posits mental states—including simple states like qualia, but also higher-order ones like beliefs—are defined by their functional role (Janet). In that way, functionalism views consciousness as a result of certain processes, or functions. From a functionalist viewpoint, an AI can be conscious, as long as it mirrors the computational aspect of consciousness performed in the human brain. This theory is seen in mathematician and computer scientist Alan Turing's famous thought experiment, the Turing Test, which he originally named the Imitation Game. In the test, a real person is placed in a text-based conversation with a computer, who is pretending to be another person. Turing claimed that if the person could not tell that the computer was not a person then the computer could be considered conscious (Turing). From a functionalist perspective, this test is sound: the computer functionally matches the human mind, and thus is indistinguishable from it. However, the computer is only following a programmatic ruleset to "win the game" and mirror the human mind, it is not truly conscious. This notion is expanded upon in philosopher John Searle's rebuttal thought experiment, the Chinese Room Argument. Searle imagines himself in a room, passing notes under the door with a man who knows Chinese. Searle is not fluent in Chinese, but he does have a Chinese-English dictionary and grammar rulebook with him that he uses to translate the notes he gets and write notes of his own back. Despite Searle's limited proficiency in Chinese, he can follow instructions to pretend he does, producing equivalent outputs to someone fluent in Chinese (Searle 3–4). In terms of functionalism, both Searle and a

real Chinese speaker functionally produce the same outputs, but their computational processes are vastly different. The Chinese Room thought experiment thus shows the core limitation of functionalism: functional equivalence to the human mind is incapable of generating consciousness, meaning a computer that looks like a human, swims like a human, and talks like a human does not necessarily think like one.

Physicalism is the school of thought that believes that consciousness is created solely by physical processes, such as electrical firing in the brain from chemical processes in the bodies of neurons. Proponents of physicalism argue that there is no mysterious non-physical component to consciousness, that it is solely due to phenomena of this world that science is yet to understand. A significant portion of physicalists believe that consciousness is directly a result of certain biological processes, which are of course held unique to the brain. One demonstration of this theory is the Dynamic Core hypothesis. This biological theory of consciousness argues that consciousness stems from the brain's "dynamic core," a clustered node of neurons located in a central part of the brain called the thalamus which integrates and modifies information. The dynamic core can modify and reorganize itself, which fosters adaptability and complex cognitive functions (Edelmann 4). The theory of the dynamic core is one of many science-based theories that try to establish a physical basis for consciousness. Most physicalist explanations, such as the dynamic core, are unique to biological systems, as they remain technologically impossible to recreate even simple neuronal behavior. However, for a physical theory to be able to discern between a conscious and non-conscious being, it would need to have fully explained the nature and origin of consciousness. That is to say, it would need to know the exact physical event that causes consciousness (and thus, a lifeform or computer would be considered conscious if it performed that event), a step no physicalist explanation has even begun to approach. This

limitation of physicalism is known as the explanatory gap, with the "gap" part referring to the difference between physical processes in the brain and conscious experience (Levine). Such a gap would need to be bridged fully before a logical assessment of whether an AI agent was conscious could be possible.

Dualism, a stark contrast to physicalism, argues that consciousness has a causal component that does not exist in the physical world, thus it cannot be explained by science. Therefore, consciousness cannot be "reduced" to pure physical activity such as neurons firing as seen in physicalist theories like the Dynamic Core hypothesis. While dualism may seem mystical or fictitious, there exist several intuitive, grounded arguments such as the "philosophical zombie" from philosopher David Chalmers. Chalmers proposes the concept of a "philosophical zombie," a person that is physically—and thus, neurologically—identical to anyone else. They walk, talk, and act like a normal person, except they lack qualia or any sort of conscious experience. Because it is theoretically and logically plausible that such a being can exist, consciousness must not be entirely supervenient (logically based) on the physical world (Chalmers 96–101). The implication is a non-physical component to consciousness which poses significant problems for the argument that an AI agent can be conscious. Because consciousness requires something non-physical, it would be impossible for a purely physical system such as a computer to truly be conscious, regardless of how well it emulates the outputs of a mind, or how complex its construction is. But there is a major limitation to dualism: it was created by humans, who, believing they were conscious, started their theory with the assumption that only humans are conscious. Because it does so, it gives an obvious bias to human minds. Dualism gives an unexplained nonphysical phenomenon as the cause of consciousness, but is unable to elaborate

further on the same phenomenon. Therefore, it is useless for analyzing the nature of consciousness in beings it hasn't already assumed are conscious themselves.

To conclude, consciousness is an unexplained phenomenon that, until recently, has only been attributed to certain biological life. Recent developments in technology have led to the idea of an expansion of this attribution to computational systems. However, the main contemporary models of consciousness, chiefly, functionalism, physicalism, and dualism are unable to assess whether or not AI agents have the capability of being conscious, because of limitations in their design and reasoning when applied to non-human (and non-human-adjacent) beings. Specifically, the functionalist narrative only accounts for output in considering whether a being is conscious and disregards the internal processing that led to that output being generated. Physicalism, while giving an adequate theory for the cause of consciousness, currently lacks the full information needed to correlate physical events in the brain to qualia, and thus cannot fully explain the cause of consciousness or allow for the determination of what is or is not conscious. Finally, dualism is fundamentally inapplicable to assessing whether a being is conscious simply due to the fact that it starts with the assumption that humans themselves are conscious, and cannot specify the nonphysical phenomena behind consciousness.

Works Cited

Block, Ned. Troubles with functionalism. 1978. Minnesota Studies in the Philosophy of Science 9:261-325.

Bory, P., Natale, S. & Katzenbach, C. “Strong and weak AI narratives: an analytical framework”. AI & Soc. 10 October 2024. https://doi.org/10.1007/s00146-024-02087-8.

Chalmers, D. J.. The conscious mind: In search of a fundamental theory. 1996. Oxford University Press.

David Bourget & David J. Chalmers. "Philosophers on Philosophy: The 2020 PhilPapers Survey". Oct 2023. Philosophers' Imprint 23 (11).

Edelman, Gerald M et al. “Biology of consciousness.” Frontiers in psychology vol. 2 4. 25 Jan. 2011, doi:10.3389/fpsyg.2011.00004

Levin, Janet, "Functionalism", The Stanford Encyclopedia of Philosophy (Summer 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.).

Levine, J. “MATERIALISM AND QUALIA: THE EXPLANATORY GAP”. Pacific Philosophical Quarterly. 1983. https://doi.org/10.1111/j.1468-0114.1983.tb00207.x

Nagel, Thomas. “What Is It Like to Be a Bat?” The Philosophical Review, vol. 83, no. 4, 1974, pp. 435–50. JSTOR, https://doi.org/10.2307/2183914.

Searle, John R. “Minds, Brains, and Programs.” Behavioral and Brain Sciences 3.3 1980. 417–424. Web.

Turing, Alan. “Computing machinery and intelligence”. Mind 59. October 1950. 433-60.