Macalester Journal of Philosophy

Volume 20 | Issue 1 Article 8

6-21-2012

Could Consciousness Emerge from a Machine Language?

Genevieve Kaess
Macalester College

Could Consciousness Emerge from a Machine Language?

Genevieve H. Kaess

Abstract Behaviorists believe the following: if the output of artificial intelligence could pass for human behavior, AI must be treated as if it produces consciousness. I will argue that this is not necessarily so. Behaviorism might be useful in the short term, since we do not know what causes consciousness, but in the long term it embodies an unnecessary hopelessness. I will attempt to establish in this essay that certain empirical knowledge of consciousness is within the realm of possibility. I will then use my own definition of certain knowledge to shed light on ways in which computer programming falls short of producing human-like consciousness.

I. Introduction

"The best reason for believing that robots might someday become conscious is that we human beings are conscious, and we are a *sort* of robot ourselves."[1] Daniel Dennett's offhand introduction to his essay "Consciousness in Human and Robot Minds" serves more generally as a summary of popular contemporary philosophical thought regarding artificial intelligence: it is possible, in theory, because human intelligence is

[1]Daniel C. Dennett, "Consciousness in human and robot minds," in *Cognition, Computation & Consciousness*, ed. Masao Io, Yasushi Miyashitatt and Edmund T. Rolls (Oxford: Oxford University Press, 1997), 17.

possible. Human life, and consciousness with it, is no more than the machinery of nature. What remains unclear is to what degree (if at all) and in what ways the mechanisms that produce human consciousness must be imitated in order to create artificial consciousness, and whether knowledge of the creation of artificial consciousness can ever be certain.

In this paper, I will argue that syntactical computer modeling is not sufficient for artificial consciousness. I will approach this point by first examining views of philosophers (specifically Alan Turing and Hilary Putnam) who have suggested behaviorism as the standard by which to judge consciousness in artificial life. I will suggest that although behaviorism provides an immediate solution to the problem of other minds, the adoption of behaviorism as a long-term solution embodies an unnecessary hopelessness regarding certain knowledge of consciousness. Applying the same standards for certain knowledge that we do for other phenomena, we can come to certain empirical knowledge of the causation of consciousness. The rejection of this claim, I will argue, is dualistic. Finally, using the standards that have traditionally been sufficient for certain knowledge, I will explain why one specific example (which I will discuss in section V) casts doubt on the claim that AI can be achieved through computer programming.

II. Definitions

For simplicity's sake, the term "artificial intelligence" (AI) will refer, in this paper, to artificial consciousness. Traditionally, consciousness has been deemed an unnecessary (or at least not necessarily necessary) condition for artificial intelligence. On the

contrary, I believe intelligence and consciousness to be inextricably linked. Intelligence, by definition, is the capacity to learn and understand[2]; understanding is a feature of consciousness. Information processing, then, can only be qualified as intelligence if it has conscious manifestation. Consciousness will be understood (in this paper) as thoughts and emotions such as humans experience them. I exclude non-human animal consciousness from my definition because the goal of AI scientists is to produce human-like intelligence (which, by my definition, entails human-like consciousness). By limiting the scope of the definition of AI in this way, a more comprehensible argument will emerge; current knowledge of the nature of consciousness in other organisms is imperfect, and any discussion of it would be based on conjecture.

The science of AI depends on the truth of one basic assumption: consciousness is a natural physical process. There is no spiritual realm of thought that exists separately from nature; therefore, provided limitless resources and a thorough understanding of the mind, we would be able to reproduce it artificially. *Computational* AI depends on the possibility that this can be realized using computer programming. In this paper, I will assume that AI is possible, but I will provide evidence that *computational* AI is not. Henceforth, "AI" will refer to computational artificial consciousness, and "computational

[2]This definitiveness of this definition is disputable. However, there is no doubt that this is one commonly used definition of "intelligence." Since I am merely using it to justify my choice to define AI in the way that I do, and not as a premise to any of my arguments, the definitiveness of my chosen definition is of little consequence.

functionalism" will be understood as the philosophical position that such AI is obtainable.

To say that the creation of artificial intelligence can be fully realized through computer programming is tantamount to saying one of two things: (1) the human mind is itself nothing more than a computer[3] – an information processing tool – or (2) computer programming and the mind can produce equivalent cognition without holding any additional features in common. I will discuss the second possibility in section VI. For the most part, computational functionalists hold that the first is true: information processing is the necessary feature of the mind. Certainly it is true that the human brain has a biological medium distinct from that of a computer, but that is all it is: a medium that realizes and supports the brain's intrinsic informational processing. Human consciousness, they believe, is a feature of the *processes*, not of the medium.

Computer programming, at its most basic level, is a series of 0s and 1s, which answer the question of whether or not various features exist. I will refer to these 0s and 1s as "computer syntax." Computer syntax is itself a mechanical feature of the computer, which is programmed in by humans. When prompted, it sets in motion a series of mechanical events within the computer that lead to the visible output on the screen or, in the case of AI, the observable actions of a robot. The 1s and 0s can be combined in very complex ways to produce impressive outcomes. In the 1950's, the research of Allen Newell and Herbert Simon suggested

[3]John R. Searle, *The Rediscovery of the Mind* (Cambridge, MA: MIT Press, 1992).

that "a computer's strings of bits could be made to stand for anything, including features of the real world, and that its programs could be used as rules for relating these features."[4]

The idea for AI was not born solely of the impressive capabilities of computers. It emerged also from the notion that computer programming is the best model for the workings of the brain. Most neurons give and receive signals in short blasts. They operate under an all or nothing principle - either they're firing or they're not. This is similar to the 1/0 duality of binary code. AI scientists posited that these neuronal impulses could be modeled by computer programming to the same effect: intelligence.

III. The Problem of Other Minds

But how would we know if that happened? Current scientific knowledge does not account for consciousness. This is called the "problem of other minds," and it is the foundation, as well as the limiting factor, for philosophical arguments regarding AI: we do not know what exactly consciousness is, and therefore we cannot test for it in others. One can only be certain of one's own consciousness. For some philosophers, this is grounds for suggesting the adoption of a behavioral standard by which we might judge what constitutes intelligence and what does not.

In his article "Computing Machinery and Intelligence," Alan Turing described his most lasting contribution to philosophy – the "Turing test." Turing devised a game in which two people (a man – "A" – and a woman – "B") sit in separate rooms as an

[4] Hubert L. Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (Cambridge, MA: MIT Press, 1997), x.

interrogator questions them. All identifying features are hidden from the interrogator. His goal is to determine which is the man and which the woman; the goal of one of the two competitors is to confuse the interrogator and the goal of the other is to help him. Turing then posed the question: "What will happen when a machine takes the part of A in this game?"[5] The interrogator now must determine which of the two is the machine. Turing asserted that if a machine could win this game as frequently as the typical human, it would be unfair to deny that it had consciousness. After all, we do not require proof of consciousness in one another. Until consciousness is de-mystified, Turing believed, we must adopt this principle of equity.

Although the Turing test is not a definitive test for consciousness, many have accepted it as the standard. We do not have the knowledge to recognize consciousness in others; therefore we are engaging in cognitive chauvinism if we suggest that a machine with humanlike cognitive capabilities (insofar as they are measurable) lacks consciousness. Turing's solution is pragmatic: to avoid prejudice, we must judge consciousness in non-humans in the same way we do in humans – behaviorally.[6] The strength of his position is that it is safe; it makes no conclusive claim about what constitutes consciousness, but instead suggests the adoption

[5] A.M. Turing, "Computing Machinery and Intelligence," *Mind* 59, no. 236 (1950): 434.

[6] I would argue that we do not always usually use behavioral characteristics to determine whether other humans are conscious. Instead, we assume that they are conscious (because of their biological status as humans) regardless of whether or not they could pass the Turing test. However, I will grant Turing this point, since it is probably true that the reason we assume humans have consciousness, even if they cannot pass a Turing test, is because as a general rule, humans behave as if they are conscious.

of a standard.

Hilary Putnam expressed slightly stronger opinions in his essay, "Robots: Machines or Artificially Created Life": we cannot expect to gain complete understanding of psychological states by studying brain physiology. "Psychological laws are only statistical … to say that a man and a robot have the same 'psychology' … is to say that the behavior of the two species is most simply and revealingly analyzed at the psychological level (in abstraction from the details of the internal physical structure), in terms of the *same* 'psychological states' and the same hypothetical parameters."[7] For example, anger is defined by one's claims and actions, not physical brain states. It is identified by behavioral features, not biological ones. This being the case, Putnam contended that "it is … necessary … that one be prepared to accept first-person statements by other members of one's linguistic community involving these predicates, at least when there is no *special* reason to distrust them."[8]

Putnam constructed the following scenario to illustrate his point: suppose that sometime in the future the robots we have invented build robots of their own (Putnam calls these "ROBOTS"). The philosopher robots then sit around debating whether or not ROBOTS have consciousness. This is akin to our current actions. Since we do not understand consciousness, we have no less duty to ascribe consciousness to robots than we do to one another. The question of consciousness, Putnam concludes, cannot currently be solved. Whether robots should be treated as if

[7] Hilary Putnam, "Robots: Machines or Artificially Created Life," *The Journal of Philosophy* 61, no. 21 (1964): 677.
[8] Putnam, *Robots*, 684.

they have consciousness, then, "calls for a decision and not for a discovery. If we are to make a decision, it seems preferable … to extend our concept so that robots *are* conscious – for 'discrimination' based on the 'softness' or 'hardness' of the body parts of a synthetic 'organism' seems as silly as discriminatory treatment of humans on the basis of skin color."[9]

The acceptance of a behavioral standard may be the most appropriate immediate solution, but Turing and Putnam seem to have been content to let it go at that. Turing declared the concept of consciousness "too meaningless to deserve discussion."[10] They adopted a perplexing stance for philosophers – agnosticism – and many contemporary philosophers are happy to follow suit; debate over consciousness is not just meaningless, they believe, but impossible to resolve. The turn to behaviorism came not from conviction of its worth, but from the lack of a better option. I will argue that such a position of hopelessness is unnecessary; consciousness can be known empirically.

The problem of other minds rests on the assumption that consciousness is accessible only through first-hand experience. But this is dualistic. If each person's consciousness exists only in a special bubble that has no physical manifestation, then it is not physical. To say that consciousness is both material in nature and fundamentally undetectable is to make a claim that is dramatically inconsistent with contemporary scientific thought. Substance is thought to break down into particles that have both charge and extension; if consciousness is material (an assumption required for

[9] Putnam, *Robots*, 691.
[10]Turing, *Computing Machinery and Intelligence*, 442.

any form of artificial intelligence), it must be detectable at some level if the detector knows where to look for it. But that is the problem: how do we figure out what to look for when we don't know what to look for? How do we make the connection between objectively viewed matter and that which we experience as consciousness?

Those who find the problem of other minds unsolvable might answer that we need proof, and that proof is impossible. First-hand experience cannot provide conclusive evidence regarding the nature of consciousness. Self-reporting is not sufficient for understanding of consciousness, because we are unaware of the causal mechanisms within our own brains. However, it seems to me that if we could thoroughly observe an individual's brain in conjunction with honest reporting of his mental states, we would discover much about the nature of consciousness, and perhaps even its causation. Honesty cannot be ensured for any given individual, but given numerous repetitions of the experiment and the assumption that most people are honest, useful data would emerge. For example, consider the following: the materialist understanding of consciousness requires that it must be possible, in theory, to replicate minds. This would be done, perhaps, by tweaking one person's neurons in various ways until the person had the personality, memories, etc. of another; the purpose of this exercise would be to learn which changes in the features of the brain are necessary for changes in consciousness.[11] Depending on how we tweaked the neurons and to what effect, we

[11] Obviously, there are ethical and practical barriers that would prevent the manifestation of this scenario, but I intend it only as a hypothetical situation to help illustrate my later point.

could draw links between brain states and conscious experience, from which we could conclusively accept or reject computational functionalism.

One objection might be that hypothetical scenarios like this one spawn sticky questions regarding personal identity. If my consciousness changes entirely to that of another person, or even if it just changes a little bit, do *I* really still exist or has my body just taken on a new identity? If I cease to exist, then clearly *I* cannot testify regarding certain knowledge of the change in my consciousness, in which case the success of the experiment (drawing links between consciousness and brain state) will depend on correct behavioral analysis. If I claim to have experienced a change from one personhood to another, in fact it suggests that I have *not* experienced such a change; upon becoming the second person, I would lose memory of the first. Even slight changes might be impervious to awareness. If I lose a memory, for example, and all memories of that memory, I cannot know that I have lost it. Self-reporting, even combined with brain observation, therefore becomes an inadequate method for the discovery of mental causation and third-person reporting of consciousness is not definitive. Furthermore, even if we *do* establish, using inductive reasoning, that a certain change in the brain produces a certain change in the *nature* of consciousness, it still does not speak to whether that feature of the brain caused that *moment* of consciousness itself. The brain might be an intermediate link in the consciousness-producing causal chain. For some philosophers, the lack of the plausibility of certain knowledge regarding the causation of consciousness is reason enough to dismiss the entire question.

Those who get caught up on the problem of other minds are forgetting one of life's early lessons: knowledge of causation in the physical world is never certain. Young children often are preoccupied by the question, "Why?" Adults who are grilled by these children are usually eventually reduced to the answer, "Because that's just the way it is." We can superficially understand causation, but when we examine our understanding, it becomes clear that all we actually do is recognize patterns. For instance, we think we understand why a ball rolls (it was pushed) and we think we understand why the push causes the ball to roll (the transference of energy). For many of us, the understanding ends there, but an expert in physics might be able to answer the question "why?" a few more times. Even our physics expert, however, is eventually forced to concede a lack of understanding. You do not wholly understand a cause if you do not understand the cause of the cause. Furthermore, all of these alleged causal understandings are actually theories based on induction. We believe that if the ball is pushed (under certain conditions), it will roll. But that belief is based on our repeated observation of this phenomenon. We have merely recognized a pattern, and concluded from it a causal relationship. Humans are only capable of identifying correlation. Causation is supposed, never known.[12]

Furthermore, we assume similarity in internal structure in entities that display similar characteristics. If a rat is born of a rat, looks like a rat and acts like a rat, we feel certain that it has internal organs much like those of other rats and we will come to

[12] David Hume, "An Enquiry Concerning Human Understanding," in *Modern Philosophy: An Anthology of Primary Sources, Second Edition*, ed. Roger Ariew and Eric Watkins (Indianapolis: Hackett Publishing Company, 2009).

conclusions based on this assumption. We believe in those conclusions with such absolute certainty that we bet our lives on them; rats are often used to test products to determine their safety for humans. If we truly believed extreme variation in the physical nature of rats possible, such tests would be worthless. Induction is by its nature uncertain, but humans trust it.

If we adopt a standard for consciousness in the name of objectivity, but refuse to accept that the causation of consciousness can be understood empirically, we have, in fact, failed to view the situation objectively. As the example of the rolling ball demonstrated, inductive correlative reasoning is good enough to use to identify other causal physical relationships. In the case of the rolling ball, we have come to the inductive conclusion that pushing the ball causes it to roll. If we repeatedly observe that a certain brain state corresponds to a certain mental characteristic, it is fair to assume causation, just as we assume that it is the push that causes a ball to roll, not that the push was an intermediate link in the causal chain[13]. Correlative evidence can demonstrate a link (or lack thereof) between brain physiology and consciousness. This evidence can be used to make conclusive claims about the nature and causation of consciousness.

Of course, the problem is that we have not yet accumulated enough correlative evidence to make conclusive claims about the causation of consciousness. But the situation is not hopeless. By adopting a position of behaviorism, one approaches this problem

[13] Additionally, if brain states are intermediate links in the causality of consciousness, then it is unlikely that syntactical modeling would produce consciousness, since it models a feature of brain states and would therefore be modeling an intermediate step.

from the wrong angle. If you turn to robots for the answer to the question of consciousness, you are looking in the wrong place. Clearly, one cannot look into a robot to determine whether or not it has consciousness. That would be like trying to determine whether something plays music without any knowledge or understanding of the nature of music. A more practical course of action is to look for the root of consciousness, and to do that, it is far wiser to look where we assume it does exist (in humans) than where we are trying to create it (robots).

IV. On Correlation

Correlation can be used in two ways. First, as I have suggested, positive correlation can lead to valid causal claims. If a light turns on every time I flip a functioning light switch, I might make the inductive claim that flipping a functional light switch causes a light to turn on. Induction is useful, but not a logically strong form of reasoning. It might be, for example, that one cause has two effects, and I correlate the two effects to each other rather than to their mutual cause. For example, a faulty light switch might produce a spark immediately after I flip it, just before the light turns on. I might induce that the spark causes the light to turn on. This would have the same inductive validity as the claim that flipping the switch turns on the light, but it would not be correct.

Negative correlation, however, is logically conclusive. Only one instance of the correlation of A and B is required to disprove the conditional statement, "If A, then not B." For example, the belief that no dogs bite humans can be disproved by the single instance of a dog biting a human. If use of computer programming to produce AI tends to have human-like results in

behavior, the behaviorist might inductively conclude that the two are equivalent and computational functionalism correct. However, it takes only one demonstration that the brain and the computer, given equivalent structural changes, produce different results to show that, at the very least, our current programming provides a flawed model of the brain.

V. Implications for Artificial Intelligence

In Section III, it was established that the search for the root of consciousness need not be futile so long as one is looking in the right place: the human brain. When we pose the question of whether AI might produce consciousness, it is important to recall that most of the initial hope for AI stemmed from its similarity to brain processes. Neurons send signals to one another with short blasts of energy, which is in some ways similar to how computers process binary code. However, it is important to note that this is not strictly true. Not all neurons fire in short bursts; some send longer signals not accounted for by computer syntax. Additionally, neurons exist in a net, whereas binary programming is linear. In his book What Computers Still Can't Do, Hubert Dreyfus described the problem of "know-how." When a person becomes an expert at a task, he no longer needs to think through all the steps of the task, but rather the proper course of action is immediately obvious. For example, a master chess player does not have to think through the rules of the game before making a move, but rather sees the position of the pieces on the board and knows instantly what to do. By contrast, the more data the computer chess player has about the game of chess, the more information it will have to analyze before making a move. Although, in general,

consciousness alone is a poor means for understanding underlying mental causation, in this case it was indicative of an underlying mechanism. Neuroscientists have explained the "know-how" phenomenon by the fact that when two neurons are simultaneously excited, the connection between them is strengthened.[14] Newer models of AI ("connectionist" models) have incorporated links like these into programming, but they are poor models for neural nets. Ultimately, even connectionist programming boils down to binary code.

For the sake of argument, however, let us grant that neuronal impulses are the source of consciousness and that binary code is a decent model for them. The question now is whether being a model is good enough to produce consciousness, or if there is some further biological feature necessary. For binary code to model neuronal information processing, one must be able to imagine that at any given moment, the neurons of the brain can be mapped syntactically. The alteration of patterns in binary code must produce output to the alteration of neuronal patterns. A recent study led by Mriganka Sur casts doubt on the causal nature of brain structure. Sur and his colleagues performed surgery on newborn ferrets,[15] so that each had one eye that sprouted connections into the part of the brain that is generally dedicated to hearing (rather than into the visual thalamus and visual cortex).

---

[14] Dreyfus, Hubert L. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, MA: MIT Press, 1997.

[15] Granted, I stated at the beginning of this paper that I was not going to tackle the notion of animal consciousness. However, the scientific community often extrapolates findings concerning animal physiology to humans, and I am assuming that this study is accurate in suggesting that there would be similar findings if we were to perform this study in humans.

There was no resulting change in the ferrets; they continued to see with the affected eyes, using the auditory portions of their brains.[16] An immediate change in neuronal patterns (and in our imaginary syntax which we have mapped onto the brain) produces no change in consciousness. This suggests plasticity of consciousness that is not observed in the output of AI. By comparison, it is difficult to believe that significant change in syntax would not produce observable change in computer function. In other words, in the case of computer syntax, there is a conditional relationship: if there is considerable change in syntax, there will be change in output.[17] For neurons, we have seen the equivalent conditional statement disproved. Here we have established lack of correlation between the result of neuronal behavior and that of syntactical programming; at the very least, we must conclude that current efforts to use computer syntax to model brain functions are fundamentally flawed. Just as a fundamental change in a recipe would not necessarily produce an observable change in outcome, but would very likely do so, this does not prove that syntax does not produce consciousness, but it suggests as much.

VI. Discussion

We have established that if neuronal impulses and syntactical programming each produce consciousness, they must

[16] Alva Noe, *Out of our Heads: Why You are Not Your Brain, and Other Lessons from the Biology of Consciousness* (New York: Hill and Wang, 2009), 53-54.
[17] One possible response to my argument would be a rejection of this claim. I am not a computer scientist, so I cannot say with absolute certainty that such a response would be unfounded. However, I think it is undisputable that if the syntax experienced the same degree of change as the neuronal impulses in this example, there would be noticeable change.

do it in different ways. Stalwart defenders of AI might claim that this is possible: that AI and the brain are fundamentally different from one another, yet produce equally valid consciousness. To defend themselves, they would likely revert to the problem of other minds. However, as I have already claimed, the problem of other minds should be dismissed as subjective. The claim that consciousness could be formed in two completely different ways is, first and foremost, unrealistic. It stems, I believe, from the belief that consciousness is spiritual – that it rises above and inhabits the physical world. If we instead accept consciousness for what it is – a biological phenomenon – it seems no more likely that computer programming (having proved dissimilar to the brain in every important way) could produce *it* than any other biological phenomenon (e.g. photosynthesis). Furthermore, if we reject the spiritual view of consciousness, yet accept that consciousness could be produced in a way that does not model the workings of the brain, we have no basis to judge what is conscious and what is not. The notion of consciousness becomes meaningless.

The conclusion to be drawn is that there is good reason to believe that syntax based AI does not produce consciousness.

Bibliography

Dennett, Daniel C. "Consciousness in human and robot minds." In *Cognition, Computation & Consciousness*, edited by Masao Io, Yasushi Miyashitatt and Edmund T. Rolls (Oxford: Oxford University Press, 1997), 17-29.

Dreyfus, Hubert L. *What Computers Still Can't Do: A Critique of Artificial Reason*. Cambridge, MA: MIT Press, 1997.

Hume, David. "An Enquiry Concerning Human Understanding." In *Modern Philosophy: An Anthology of Primary Sources, Second Edition*, edited by Roger Ariew and Eric Watkins (Indianapolis: Hackett Publishing Company, 2009), 533-600.

Noë, Alva. *Out of our Heads: Why You are Not Your Brain, and Other Lessons from the Biology of Consciousness*. New York: Hill and Wang, 2009.

Putnam, Hilary. "Robots: Machines or Artificially Created Life." *The Journal of Philosophy* 61, no. 21 (1964): 668-691. <http://www.jstor.org/stable/2023045>

Searle, John R. *The Rediscovery of the Mind.* Cambridge, MA: MIT Press. 1992.

Turing, A.M. "Computing Machinery and Intelligence." *Mind* 59, no.236 (1950): 433-460. <http://www.jstor.org/stable/2251299>