2019

"Alexa, Are You Conscious?": Exploring the Possibility of Machine Consciousness

Emma Cornwell

“ALEXA, ARE YOU CONSCIOUS?”: EXPLORING THE POSSIBILITY OF MACHINE CONSCIOUSNESS

by

EMMA CORNWELL

SUBMITTED TO SCRIPPS COLLEGE IN PARTIAL FULFILLMENT OF THE DEGREE OF BACHELOR OF ARTS

PROFESSOR SCOTT-KAKURES
PROFESSOR KIND

APRIL 26, 2019

INTRODUCTION

The notion of a conscious machine is by no means a new one. Science-fiction stories of computers and robots gaining consciousness infiltrate nearly every medium of today's pop culture. But these stories are just that: stories. And while the glorified narratives that we see and hear about of computers falling in love or anthropomorphic robots taking over the world are certainly exaggerated, the basic notion behind them is not. Significant progress has undoubtedly been made recently in the field of artificial intelligence. This progress has been so significant even that the idea that a truly conscious machine may exist (perhaps even someday soon) is not far-fetched or unheard of. But is this actually possible? What would machine consciousness look like?

This thesis seeks to answer the following question: "could a machine be capable of consciousness?" I begin to tackle this question by providing a presumed definition of consciousness, employing Bernard Baars' Global Workspace Theory. Next, I look to various discussions of machine intelligence and whether or not this would be sufficient for categorizing a machine as conscious. And lastly, I explore the notion that the human brain may be a sort of computational system itself and the implications this notion has for the potential that non-human systems may achieve consciousness.

Through these sections, I ultimately conclude that a machine could potentially mimic the cognitive systems of the human brain that produce consciousness (at least insofar as these systems and consciousness itself are defined by Global Workspace Theory). And therefore, a machine could indeed be capable of consciousness.

SECTION 1:

PRESUMING A DEFINITION OF CONSCIOUSNESS

In order to determine at what point a machine might become conscious, I must first provide a presumed definition of consciousness to which I will refer throughout this thesis. For the purpose of this argument, I will employ Bernard Baars' Global Workspace Theory, which he details in his 1988 book entitled *A Cognitive Theory of Consciousness*. While a complete examination of this book and its explanation of Global Workspace Theory is beyond the scope of this thesis, I will provide a broad overview which is necessary to my argument. First, I will outline the models Baars develops and the general path he takes in *A Cognitive Theory of Consciousness* to explain how this definition of consciousness actually functions (Baars, 1988).

Baars initially discovers a way in which to state empirical constraints which any adequate theory of consciousness would meet: using multiple contrasting pairs of similar events in which one event, given any adequate theory of consciousness, would be considered conscious while the other event would not be (i.e. a conscious image of one's morning breakfast in contrast with the same image of the same breakfast but in one's memory and unconscious) (Baars, 1988, 1.2.2).

From this, Baars asserts that there is a system architecture, called the "global workspace," in which internally consistent conscious representations are "globally

broadcast" to an assortment of specialized unconscious processors (Baars, 1988, Model 1).

He then provides the neural basis for conscious experience by consulting neurophysiological evidence (via the Extended Reticular-Thalamic Activating System), determining it to be largely consistent with the aforementioned framework (with the exception of the fact that there is evidence not solely for global broadcasting, but moreover, there is evidence for feedback from the receivers of the global messages and also in return to the source of the message) (Baars, 1988, Chapter 3).

He then notes that some unconscious networks, referred to as "contexts," shape conscious contents. Contexts can work collaboratively to cooperatively constrain conscious events. It is at this point that Baars determines that both global broadcasting and internal consistency are necessary conditions for conscious experience (Baars, 1988, Model 2).

Then, to explain why repeated events tend to fade out of consciousness while still being processed unconsciously, Baars introduces another necessary condition for consciousness: informativeness. He defines informativeness as "a choice within a context of alternatives, demanding adaptation by other processors;" therefore, this necessary condition generates an operative function for the feedback coming from the receiving processors to the global workspace. He asserts that conscious experience requires some level of adaptation and that conscious events must be new or noteworthy in order to trigger adaptive processing (Baars, 1988, Model 3).

Next, he moves onto the concept of intentions, which are mostly unconscious goal structures that utilize conscious goal images to employ subgoals to achieve their goals. Notably, it is at this point that he suggests the idea of the stream of consciousness: intermingling conscious events and goal contexts (Baars, 1988, Model 4).

From here, Baars turns to William James' ideomotor theory of voluntary control to give an interpretation which proposes that, in the absence of any conflicting goal images or intentions, conscious goal images alone are able to prompt actions. Therefore, it follows that the conscious aspect of voluntary action is covertly shaped by several unconscious conditions. Here, Baars is finally capable of making sense of the association between actual qualitative conscious events (i.e. perception and imagery) and the existence of nonqualitative conscious contents (i.e. concepts and instantaneous intentions to act). After positing that there are momentary conscious images at work in abstract concepts and intentions (although these images are tough to recover), he concludes that "all conscious events involve qualitative phenomena, even though some may be quite difficult to retrieve" (Baars, 1988, Model 5).

Baars then offers that there is a difference between "conscious experience as a subjectively passive state" and "attention as the active control of access to consciousness." The distinction between these two seamlessly fits into Global Workspace Theory (Baars, 1988, Model 6).

He then shifts to deciphering the ideas of the "self" and the "self-concept." Using the same minimal contrasts method, he offers a better understanding of the

notion of the "self." He suggests that the "self" can be defined as the thing undergoing the context of experience and which "serves to organize and stabilize experiences across many different local contexts," while the "self-concept" can be defined as "a control system that makes use of consciousness to monitor, evaluate, and control the self-system" (Baars, 1988, Model 7).

Lastly, given all of these developments, Baars concludes that conscious experience plays several important roles in the nervous system. He lists 18 basic adaptive functions of consciousness (see Baars, 1988, Chapter 10), the most essential of which is consciousness' capability of aiding in cooperative interaction between several sources of knowledge in order to deal with newness (Baars, 1988, Chapter 10).

Finally, via the progression of these models, Baars arrives at 5 necessary conditions for conscious experience which will be vital to the definition of consciousness to which I will be referring in this thesis. They are as follows (Baars, 1988):

1. "Conscious events involve globally broadcast information"
2. "Conscious events are internally consistent"
3. "Conscious events are informative- that is, they place a demand for adaptation on other parts of the system"
4. "Conscious events require access by a self-system"
5. "Conscious experience may require perceptual or imaginal events lasting for some minimum duration"

Before concluding this outline, I would like to address 2 replies to Baars' Global Workspace Theory which merit acknowledgement in order to determine whether or not this is a fully sufficient definition of consciousness for the purpose of this thesis.

The first of these replies comes from Ned Block's "On a Confusion About a Function of Consciousness." Block asserts that consciousness is a "mongrel" (hybrid) concept and that the word consciousness itself "connotes a number of different concepts and denotes a number of different phenomena" (Block, 1995). In other words, Block is suggesting that there is more than one thing to which "consciousness" refers.

Block argues that Global Workspace Theory incorrectly conflates phenomenal consciousness with access consciousness (two uniquely different types of consciousness), suggesting that Baars continuously asserts that he is discussing phenomenal consciousness despite the Global Workspace Theory actually being a model of access consciousness (Block, 1995).

If we accept this argument and the classification of consciousness as a "mongrel" concept, then Global Workspace Theory is indeed not a viable functioning definition of consciousness for this thesis because it would be unclear if the sort of consciousness present in humans is indeed the same sort of consciousness that could possibly be present in machines.

However, I will instead accept a direct reply to Block from Baars himself in a paper entitled "Evidence that Phenomenal Consciousness is the Same as Access Consciousness," so that we can dismiss Block's reply and continue to employ

Global Workspace Theory as the functioning definition of consciousness for this thesis.

In this rebuttal, Baars disputes Block's assertion that consciousness is a "mongrel" concept in favor of the notion that consciousness is instead "a single core fact with many facets" (Baars, 1995). He also points out that this distinction is, in fact, an empirical one for which Block gives no evidence. And so, citing his own initial research, he reasserts that consciousness is indeed a "unified problem with many superficially different aspects," and thus we may actually equate phenomenal consciousness and access consciousness. (Baars, 1995, 1988). Therefore, Global Workspace Theory remains a sustainable definition of consciousness for the purpose of this thesis.

Another reply to Global Workspace Theory which ought to be acknowledged in order to deem this definition of consciousness viable is Susan Blackmore's challenge of the concept of stream of consciousness in her paper entitled "There is No Stream of Consciousness."

In this paper, Blackmore claims that consciousness is a "grand illusion" (Blackmore, 2002). This assertion merits recognition in this thesis because if we accept that consciousness is merely an illusion and that these spaces which Baars refers to do not exist, then any sort of consciousness a machine may possess might also be illusory. In this case, we would have no way of knowing if a machine had actually reached consciousness or if the machine was merely purporting itself to be conscious, and this thesis does not ask the question: "can a machine possess the illusion of consciousness?"

However, Blackwell's claim and Global Workspace Theory do not seem to be mutually exclusive. While the notion of the stream of consciousness which Blackwell points to certainly could be illusory, this illusion may in fact simply be a byproduct of the processes at work in Global Workspace Theory. Therefore, whether or not stream of consciousness is an illusion is irrelevant to concluding if a being and/or machine is conscious. And thus, we can maintain that Global Workspace Theory is indeed a feasible presumed definition of consciousness for my arguments.

SECTION 2:

MACHINE INTELLEGENCE

Given the presumed definition of consciousness (Global Workspace Theory) which I outlined in the previous section, I will now move onto examining the role intelligence plays in determining whether or not a machine can be said to have reached consciousness.

The Turing test, proposed by Alan Turing in "Computing Machinery and Intelligence", assesses machine intelligence (Turing, 1950). Using the test, we can attempt to decide whether or not a machine's behavior is indistinguishable from a human's behavior. The machine is assessed by a human evaluator who holds conversations with many participants through conversational text on a computer. All of these participants are humans except the machine which is being evaluated. The machine will have passed the Turing test if the human evaluating the machine ultimately cannot distinguish between the human participants and the machine participant. Turing concludes that if a machine can pass the test, it is then capable of intelligence (Turing, 1950).

While this process is now commonly referred to as simply the Turing test, Alan Turing initially called it "the imitation game" (Turing, 1950). The imitation game works as follows: a neutral human interrogator resides in a room separate from two other human participants, one of which is a female and one of which is a male (A and B, respectively). The interrogator labels the two beings X and Y. The

interrogator then asks a series of questions in an attempt to decipher which participant is X and which participant is Y. In order to mask the participants' voices, all answers are given through a typed medium. Both participants, however, are meant to attempt to convince the interrogator that they are the female participant.

Turing then expands this thought experiment of the imitation game to machine intelligence by replacing participant A with a computer (Turing, 1950). Turing begins with the imitation game and eventually replaces one of the participants with a machine to show that only beings that are capable of intelligence (such as humans) are capable of passing the test. And thus, if a machine could act in the way that these humans could, we would consider it to be intelligent as well (Turing, 1950).

Robert French, though, argues that the Turing test is not a viable measurement of intelligence, since it tests only for intelligence which would be present in humans (French, 1990). He concludes that "the [Turing] test provides a guarantee not of intelligence but of culturally-oriented human intelligence" (French, 1990, 54). To combat this, he constructs what he calls the seagull test, which he then compares to the Turing test.

The seagull test is a theoretical test for flight. The test goes that a machine can be considered capable of flight if a neutral onlooker cannot distinguish between the machine and a real seagull when looking at a three-dimensional radar screen (French, 1990). Of course, any machine which passes the seagull test would indeed be capable of flight, but there are undoubtedly a number of flight-

capable things which would *not* pass the test since the test's expectations of flight capability are much too high and much too specific (French, 1990). For example, an airplane would not be indistinguishable from a seagull on a three-dimensional radar screen, but an airplane seems capable of some sort of flight.

French concludes then, that since the capacity for flight should not be equated to a seagull's version of what flight is, intelligence should not be equated to a human's version of what intelligence is. This, he argues, is how the Turing test fails (French, 1990).

We must, therefore, consider the notion that being human may not necessary for the sort of intelligence posited by the Turing test and that this "culturally-oriented human intelligence" may not be the only sort of intelligence to exist (French, 1990, 54).

Human brains function much in the same way that machines do: we receive an input which is processed through our biological "program" via our brains "electrical wires," and then we give an output. We perceive this process as intelligence, and, as humans, we would quite likely pass the Turing test (unless, of course, we were intentionally trying to fail and deceive the interrogator) given the fact that the test itself assesses whether or not a computer is indistinguishable from a human.

If we take this view to be true and concede that our brains are *also* just computers following a program, yet we still consider ourselves to have thought and understanding in addition to intelligence, then I conclude that a machine which is capable of passing the Turing test could potentially have the capacity for

thought and understanding in addition to being capable of intelligence and furthermore that being biologically human is not necessary for thought and understanding.

Alan Turing himself addresses several criticisms of his argument. Importantly, though, is the objection which he calls the argument from consciousness (Turing, 1950, 446). The argument from consciousness says that the Turing test is not a viable test of machine intelligence because it does not test for consciousness and only conscious beings are capable of intelligence. This objection is simply a "denial of the validity of [the] test," Turing says, and takes on a solipsist mindset which claims that "the only way by which one could be sure that a machine thinks is to *be* the machine and to feel oneself thinking" (Turing, 1950, 446, emphasis in original).

Turing combats this suggestion by simply claiming that anyone who agrees with the argument from consciousness could eventually be convinced to abandon their claims, although he gives no further explanation for this (Turing, 1950, 446). This seems to be a weak and insufficient response, as it is dismissive and does not give any further explanation as to *why* these people could be convinced otherwise.

Furthermore, Turing acknowledges that "[he] do[es] not wish to give the impression that [he] think[s] there is no mystery about consciousness;" however, he also emphasizes that he "do[es] not think these mysteries necessarily need to be solved before we can answer the question with which we are concerned in this

paper" (Turing, 1950, 447). He therefore does not believe that consciousness is required for intelligence.

Turing indeed understands that his test is insufficient for determining machine consciousness, he just does not feel a need to address the notion of machine consciousness at all (Turing, 1950). His interests, at least in this paper, lie in machine thinking and machine intelligence, not in machine consciousness, as made clear by his assertion that he does not need to address the problem of machine consciousness (Turing, 1950).

Turing's implication, then, is that there could in fact be intelligent machines which are not also conscious machines. This implication is essential to my assertion that intelligence is a necessary condition for determining if a machine is conscious. I maintain that even if a machine could truly possess the ability to think and ultimately to be categorized as intelligent, this is not a sufficient condition for considering the machine to be conscious. It is, however, a necessary one. A system, no matter what its biological makeup is, must be thinking, at least to some extent, to be conscious, given the fact that any sort of process at work in Global Workspace Theory is a thinking one (in the sense that they process information and function as a part of a system).

Another critic of the Turing test, John Searle, objects to the notion of strong AI (which suggests that machines may actually *be* intelligent and capable of actual thought and understanding as we conceive of it) in his 1987 paper entitled "Minds, Brains, and Programs." In this paper, Searle argues for the view of weak AI, which proposes that machines can merely simulate intelligence

(Searle, 1987). He claims that no matter how advanced an AI becomes, it will never and indeed can never truly *understand* its outputs. The machine may appear human, but Searle holds that no matter how much a machine's behavior mimics human behavior, it will never actually possess authentic intelligence and/or understanding. He argues for this view by proposing the Chinese Room thought experiment (Searle, 1987).

In this thought experiment, Searle theorizes that he is in a room with ample writing materials as well as an English rule book for translating Chinese characters. Searle is given a task: he is fed inputs of Chinese characters, he utilizes the English rule book, and then he writes an output which is also in Chinese characters. Searle claims that he too, in this room, would pass the Turing test, despite never fully understanding Chinese since he is merely using a rule book which is written in his native English to decipher his Chinese inputs and construct his Chinese outputs (Searle, 1987).

Ultimately, Searle asserts that, like himself, the machine would not have a genuine understanding of Chinese. However, many proponents of strong AI have replied to Searle's Chinese room experiment, exposing its flaws.

Perhaps the most convincing reply to Searle's Chinese room experiment is what is known as the robot reply. The robot reply proposes that we can imagine a robot with a computer for its brain which would not only receive symbols as inputs and construct them as outputs, but which would also be able to perform human-like motor functions (i.e. it would have wheels for getting around, a camera for seeing, arms for moving objects, etc.). Therefore, this robot could

perform bodily actions via its computer brain much in the same way a human does. This robot, then would be able to learn through its physical being, as it would be able to utilize its sight and its physical actions to acquire knowledge and improve its computer brain, just as humans do as they grow up and learn more about the world around them (Searle, 1987, 420).

However, vital to Searle's refutation of the robot reply is the notion that the Chinese symbols do not actually have any *real* meaning. And since they do not have *real* meaning, they cannot yield *real* understanding. Therefore, the robot in the Chinese room could not, in fact, understand. Searle further refutes the robot reply by supposing that instead of putting a computer in the robot, he would make himself what he calls the robot's homunculus. Searle still does not understand Chinese, but he can perform the task just as well as the robot's computer brain by following his rule book. All Searle *truly* understands, he argues, is symbol manipulation (Searle 1987, 420).

Searle goes on to claim that "the robot has no intentional states at all; it is simply moving about as a result of its electrical wiring and its program" (Searle, 1987, 420). Moreover, when Searle is functioning as the robot's homunculus, he insists that he too has no relevant intentional states since he is merely following the instructions he has been given.

He insists that instantiating the same program is not a sufficient condition for having propositional attitudes such as humans have since this instantiation is merely the robot following its program (Searle, 1987, 420). Therefore, Searle argues, neither the computer brain nor Searle himself, when he is functioning as

the robot's homunculus, actually *understands* Chinese since they are both mere instantiations of a program and are simply following what they are have been instructed to do. Ultimately, Searle concludes that both systems (the robot and Searle himself) have no intentionality (Searle, 1987, 420).

However, in "Searle on What Only Brains Can Do," Jerry Fodor suggests that Searle's refutation of the robot reply remains unsatisfactory. Fodor argues that robots, in fact, do not merely instantiate programs, as Searle has claimed, since they are able to physically interact with their surroundings and even gain further knowledge in order to perform better by learning from these interactions through their experiences, much in the same manner that humans learn about the world as they grow up (Fodor, 1980). He notes that Searle fails to provide us with any sort of reason for believing that "biochemistry is important for intentionality and, prima facie, the idea that what counts is how the organism is connected to the world seems far more plausible" (Fodor, 1980, 431).

Fodor argues for a causal theory: that if a system (which indeed is not necessarily a human system, for it could be a machine system, an animal system, etc.) possesses *both* a program for manipulating symbols and a way of interacting with the world, then those symbols surely have meaning because they provide a function and a purpose (this is opposed to Searle's objection to the robot reply in which the Chinese symbols actually have no meaning) (Fodor, 1980). And if symbols have meaning, it follows that they have representational contents, meaning they represent an internal brain-state to the external world. Therefore, they have intentionality, since they possess both of these properties. In the case of

Searle's refutation of the robot reply, both the computer brain and Searle himself functioning as the robot's homunculus, then, have intentionality (Fodor, 1980, 431).

Searle's refutation of the robot reply may indeed hold if we take human brains to be more complex and more extraordinary than the computers which we engineer ourselves, since that would suppose that the human brain is not simply a computer program. However, I argue that the human brain, while certainly complex, is not that special.

I agree with Fodor's argument that biology is not important for considering intentionality because a computer program may be able to possess intentional states and because human brains are not necessarily special in terms of their abilities (they may just be computer programs themselves) (Fodor, 1980). Therefore, a computer could in fact be capable of possessing propositional attitudes if it possessed "the right kind of causal relation[ship]" to the world around it, one that was capable of intentionality (Fodor, 1980, 431). I argue that perhaps we as humans are *also* "simply moving about as a result of [our] electrical wiring and [our] program" (Searle, 1980, 420). Despite the difference in the actual biological materials which make up a human brain and the manufactured materials which make up a nonhuman computer, the two systems may function very much in the same way.

Ultimately, Searle fails to adequately defend his Chinese room thought experiment since he rejects this notion that human brains could actually be operating identically to the way computers operate. Fodor asserts that as long as

the computer or machine possesses a program for manipulating symbols and a way of interacting with the world, we can conclude that it has intentionality (representational properties of internal brain-states to the outside world) (Fodor, 1980).

Furthermore, humans also have physical bodies which allow us to interact with the world around us, helping us learn as we have more and more of these interactions. Therefore, human beings also satisfy Fodor's conditions for intentionality. But it may not be our biological makeup which allows us to be deemed capable of thought and understanding, since it is not necessary that the program for manipulating symbols be biologically composed of a human brain or that the way of interacting with the world be composed of a human body. Therefore, a system need not be human to be capable of thought and understanding, and thus we may conceive of a machine which could be capable of such as well.

If a machine were to be created which was ultimately able to *fully* pass the Turing test and which *also* possessed both of Fodor's conditions for intentionality (again, there are that the system has 1. a program for manipulating symbols and 2. a way of interacting with the world) (Fodor, 1980), then I argue that we may conclude that this machine would function identically to the ways in which human beings function. Furthermore, this machine would not be merely simulating these same sorts of functions, but rather it would be *genuinely* functioning in the same way that humans do (and therefore, it could possess the possibility for being conscious, but I will address this conclusion further in Section 3). We could

conclude, then that this intelligent machine would possess propositional attitudes and intentionality. And therefore, it would ultimately be truly capable of thought and understanding.

Ultimately, a machine which could both pass the Turing test *and* possess Fodor's conditions for intentionality could be *genuinely* capable of thought and understanding since it would be both capable of intelligence and possess intentional (representational) properties. But this conclusion has important implications for considering whether or not machines which eventually (or even which hypothetically) reach this level of sophistication and complexity ought to be considered conscious beings in accordance with Global Workspace Theory. While we may conclude that these machines may be thinking and understanding and that they should be considered to be intelligent, this may not be a sufficient condition for their having consciousness, which is comprised of more than just a capacity for intelligence.

SECTION 3:

HUMAN BRAINS AS COMPUTATIONAL

Now, I will expand upon the notion of the human brain possessing computational properties, which I briefly touched on in the previous section and which plays an important role in determining whether or not a machine could be capable of consciousness.

The notion that our human minds themselves can be categorized as machines is not a new one. In her book entitled *Mind as Machine: A History of Cognitive Science*, Margaret Boden proposes that the idea that we may comprehend the mind as a kind of machine is a crucial and unique feature of cognitive science (Boden, 2006). For Boden, cognitive science is not restricted to traditional definitions of "cognition" (which focus on concepts such as knowledge and reasoning. Rather, she expands the classification of "cognition" to include notions like personality, emotion, communication, sociality, and action (Boden, 2006).

If we accept Boden's notion of the mind as a machine, we may have a better way of understanding how other sorts of machines may function similarly to the human mind. While an exploration of the history cognitive science and a complete evaluation of Boden's book is beyond the scope of this thesis, Boden's suggestion lays the foundation for thinking about how a machine could think and/or be conscious.

While Boden's book references the human *mind* as a machine, Patricia and Paul Churchland categorize the human *brain* as "a kind of computer" in their 1990 article entitled "Could a Machine Think?" (Churchland and Churchland, 1990, 37). Although the human mind and the human brain are two distinct entities, they are both relevant to comparing human consciousness to machine consciousness. Churchland and Churchland's article touches on a similar notion to Boden's book, while further suggesting that systems which mimic the human brain may indeed generate a truly conscious machine. (Churchland and Churchland, 1990).

Furthermore, Churchland and Churchland agree with John Searle's rejection of the notion that the Turing test is a sufficient condition for conscious intelligence (Churchland and Churchland, 1990). They also note that, to some degree, their reasons for rejecting this notion are similar to those of Searle: both agree that the means by which the input-output function is realized is imperative. In their words, both Churchland and Churchland and Searle think that "it is important that the right sorts of things be going on inside the artificial machine" (Churchland and Churchland, 1990, 37).

However, while Searle's argument stems from commonsensical instincts regarding whether or not semantic content is present, Churchland and Churchland ground their argument "on the specific behavioral failures of the classical SM machines and on the specific virtues of machines with a more brainlike architecture" and ultimately conclude that the brain "need not be the only physical

system capable of [making systematic use of computational advantages which the brain does]" (Churchland and Churchland, 1990, 37).

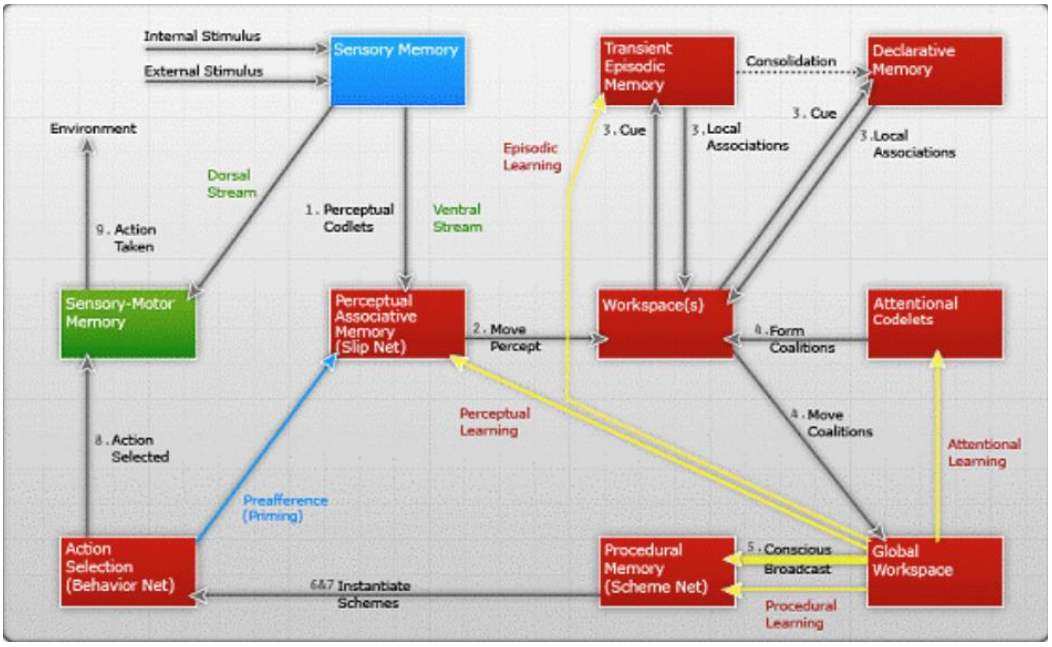
Ultimately, I agree with Churchland and Churchland's conclusion that a system which mimics a human brain (if we accept the human to be conscious) could potentially possess consciousness itself. As I have asserted before, we have no reason to believe that the human brain is special in its abilities. Thus, given these arguments, an AI which theoretically does not possess the same biological components as a human brain but which parallels the same sort of system could be a persuasive prospect for machine consciousness (again, with consciousness being defined via Global Workspace Theory).

Finally, I will conclude by introducing a final paper by Bernard Baars and Stan Franklin entitled "Consciousness is Computational: The LIDA Model of Global Workspace Theory." I have chosen to end with this paper because it proposes a viable application of Global Workspace Theory to the possibility of machine consciousness.

Baars and Franklin begin with the notion that any understanding of machine consciousness will be contingent on an understanding of human consciousness. For this purpose, they rely on Global Workspace Theory, given its consideration of the differing conscious and unconscious aspects of human cognition and its accounting for the correspondence of the vastly unconscious processing of the human brain and the fleeting capacity for conscious stream of thought (Baars and Franklin, 2009).

The LIDA model, principally based on Global Workspace Theory, is “a comprehensive, conceptual and computational model covering a large portion of human cognition” (Baars and Franklin, 2009, 3). However, the model, grounded in its cognitive cycle, can be applied to “every autonomous agent, be it human, animal, or artificial” (Baars and Franklin, 2009, 3)

A further explanation of the LIDA model can be found in “Consciousness is Computational: The LIDA Model of Global Workspace Theory,” Baars and Franklin, 2009, Section 3. A figure of the LIDA cognitive cycle can be found below (Baars and Franklin, 2009, Figure 1):



Baars and Franklin ultimately propose, as many others (including myself) have, that the human brain functions much in the same way that a computer does. Their claim, though, is evidenced by the LIDA cognitive cycle. They conclude

that “the conscious (and well as the non-conscious) aspects of human thinking, planning and perception and produced by adaptive, biological algorithms” (Baars and Franklin, 2009, 9). In other words, in humans, functions such as thinking, planning, and perception are merely complex collections of processes. These processes make up a sort of computational program itself.

Given this, Baars and Franklin also suggest that “machine consciousness may be produced by similar adaptive algorithms running on the machine” (Baars and Franklin, 2009, 9). That is to say, if a machine’s algorithms were to mimic those of a human brain, we would have reason to believe that a machine could be conscious as well.

Given Boden’s broader classification of “cognition,” Churchland and Churchland’s conclusion that a system parallel to that of the human mind could conceivably conscious, and Baars and Franklin’s notion that algorithms of human consciousness could be mimicked to produce machine consciousness, I conclude that human brains, minds, cognition, and overall consciousness are merely programs themselves. And thus, a machine could potentially mimic these systems in order to reach the same sort of consciousness as defined by Global Workspace Theory.

CONCLUSION

This thesis pursued the following question: “could a machine be capable of consciousness?” In Section 1, I offered a presumed definition of consciousness for the purpose of my argument, utilizing Bernard Baars’ Global Workspace Theory. In Section 2, I discussed several considerations of machine intelligence and contemplated if they would be sufficient for categorizing a machine as conscious. Finally, in Section 3, I turned to the idea that the human brain may function as a computational system itself and concluded that given this idea, it is possible that non-human systems may achieve consciousness if they were to mimic human systems which give way to consciousness (at least insofar as these systems and consciousness itself are defined by Global Workspace Theory). Considering this, I conclude that the answer to my proposed question is yes, a machine could indeed be capable of consciousness.

WORKS CITED

Baars, Bernard J. *A Cognitive Theory of Consciousness.* Cambridge University Press, Cambridge, 1988.

Baars, Bernard J. "Evidence that Phenomenal Consciousness is the Same as Access Consciousness." *Behavioral and Brain Sciences*, vol. 18, no. 2, 1995, pp. 249, http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/1995_Function.pdf.

Baars, Bernard J., and Stan Franklin. "Consciousness is Computational: The LIDA Model of Global Workspace theory." *International Journal of Machine Consciousness*, 2009, https://pdfs.semanticscholar.org/0162/9fcc96dd70002569a1d17ec3f13348323680.pdf.

Blackmore, Susan. "There is no Stream of Consciousness." *Journal of Consciousness Studies*, vol. 9, no. 5, 2002, pp. 17-28, https://www.susanblackmore.uk/articles/there-is-no-stream-of-consciousness/.

Block, Ned. "On a Confusion about a Function of Consciousness." *Behavioral and Brain Sciences*, vol. 18, no. 2, 1995, pp. 227–

247, http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/1995_Function.pdf.

Boden, Margaret. *Mind as Machine: A History of Cognitive Science.* Oxford University Press, 2006.

Churchland, Patricia, and Churchland, Paul. "Could a Machine Think?" *Scientific American*, 1990, pp. 32-27, https://www.scientificamerican.com/index.cfm/_api/render/file/?method=inline&fileID=9DDB5D39-D5BC-4460-845FDB846BA6FCA8.

Fodor, Jerry A. "Searle on what Only Brains can Do." *Behavioral and Brain Sciences*, vol. 3, no. 3, 1980, pp. 431. https://www-cambridge-org.ccl.idm.oclc.org/core/journals/behavioral-and-brain-sciences/article/searle-on-what-only-brains-can-do/B5DE896E6DBB9CB8879D4E8B231D9312.

French, Robert M. "Subcognition and the Limits of the Turing Test." *Mind: A Quarterly Review of Philosophy*, vol. 99, no. 393, 1990, pp. 53-65, http://search.ebscohost.com/login.aspx?direct=true&db=phl&AN=PHL1778875&site=ehost-live&scope=site.

Searle, John R. "Minds, Brains and Programs." St Martin's Press, New York, 1987. https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/minds-brains-and-programs/DC644B47A4299C637C89772FACC2706A.

Turing, Alan M. "Computing Machinery and Intelligence." *Mind: A Quarterly Review of Philosophy*, vol. 59, 1950, pp. 433-460, http://search.ebscohost.com/login.aspx?direct=true&db=phl&AN=PHL1061189&site=ehost-live&scope=site.