HIGH SCHOOL EDITION
Journal of Student Research
Volume 12 Issue 2 (2023)

AI and the Neurobiology of Consciousness

Chandni Kumar[1] and Tom McClelland[#]

[1]Bellaire Senior High School
[#]Advisor

ABSTRACT

Consciousness has confounded patrons of both philosophy and hard sciences for centuries. With new technologies coming into the market and increased usage of medical treatments involving the alteration of brain signals, an extra layer of complexity is added to the topic: to what extent does AI affect the neurobiology of human consciousness? Using the Hypothesis of Extended Cognition and Information Integration Theory, this paper will explore the impacts of AI on consciousness.

Introduction

Consciousness has been a notoriously elusive topic in both hard sciences and philosophy throughout history. As science and philosophy have progressed to be able to disprove certain theories about consciousness, we have developed fairly decent criteria that one has to meet to be conscious-namely wakefulness and awareness. This will be relevant once the validity of functionalism is established in relation to the topic. As times change and technologies develop, we add more variables to the already vague understanding of consciousness. With neural implants such as Neuralink coming into the market and increased usage of Deep Brain Stimulation (DBS) to treat neurological illness, it is important to examine the impact it has on our conscious mind and the possibility of extended consciousness[1].

Dualism vs. Physicalism

I believe it is necessary to adopt a functionalist view to understand consciousness but to do this one must distinguish physicalism from dualism. I disagree with dualism, but enough acknowledge it as valid to warrant an argument as to why dualism is a futile approach concerning this topic. There is no dualist perspective that can't come up with some way around logic and reason by endorsing the idea of a mind that is separate from the brain yet somehow still important. The greatest offender in this is the epiphenomenalist, who believes that the mind is separate from the brain and mental events have no impact on physical events, but physical events cause mental events. It's a brilliant circle for the sake of continuing an argument, but it makes it impossible to actually solve anything because the epiphenomenalist can simply say that X (physical event) caused Y (mental event), but X also caused Z (physical event) which occurred directly after Y. This raises an issue because Y is disregarded as a cause for Z for no reason other than it occurred in the 'mind'. There is no discernible way to combat this as every 'mental' state is written off as occurring in the brain, which itself would not raise concern if it weren't insisted that there was a mind separate from the brain. The mind is described as a vestigial organ, yet it still has a purpose, but there is no event one can describe as inherently 'mental' that doesn't have some sort of physical impact. If someone is experiencing sadness (something classified as mental), they are likely to look

[1] Extended consciousness can be described as an awareness of surroundings not in one's immediate environment or consciousness of inanimate components such as BlueTooth

upset or cry. It would seem that the mental state of sadness is the reason for the physical reaction of crying, but the epiphenomenalist will say that crying is a physical event caused by the original event with no relation to the mental state. My qualm with this is that a physical event caused the mental event, but the inciting 'sad event' wouldn't make us cry if it weren't processed as a mental event. Why would we cry or get upset at a situation that didn't make us sad or have some sort of mental alteration? It is these fallacies that point me towards physicalism, the idea that mental events are physical events and everything can be explained physically even if we don't have the information yet.

Functionalism's Validity

Functionalism is a form of physicalism/materialism subscribing to the idea "that what makes something a mental state of a particular type does not depend on its internal constitution, but rather on the way it functions, or the role it plays, in the system of which it is a part (Levin, 2018)." This means that the internal process does not have to be the same for every human, but that the function of these processes determines consciousness. A less compelling idea under the physicalist umbrella is identity theory, the idea that all events are defined by one physical process and any alternative path renders the human as non-conscious. This criterion is simply too strict, as considering our rather limited understanding of consciousness, it would be unwise to rule out a being as conscious. If the Identity Theorist's argument does not seem unreasonable, we can use the example of neuroplasticity and childhood brain trauma. If a child had a partial lobectomy of the occipital lobe as a three-year-old, the plasticity of a young brain would accommodate this during development and the child would eventually be able to recognize faces and see normally, even if the blind spot is a bit larger. One cannot logically say that this child is not a conscious human being because they have facial recognition processes through a different network than one with a 'typical' brain. If using the brain as an example is too much of a grey area we can look at real valve replacement procedures on the human heart; a synthetic valve replaces the original valve's function for the person to stay alive, but it simply wouldn't make sense for someone to say that the organ while serving the purpose, is no longer a heart because it has a small plastic element. One may raise the issue that the heart is not the brain so the point is null, but if Identity Theory dictates that in order for someone to have 'human consciousness' they must have all typical structures of a human, a conscious human cannot have any anatomical anomalies. Even if an Identity Theorist were to say the idea is specified to the brain, would the blood passing through the valve into the brain not invalidate that conclusion?

The Neurobiology of Consciousness

Now arises the question: how does the merging of AI and the human brain impact consciousness? To understand the impact of alterations to conscious experience, we must gather information to estimate the workings of natural consciousness. By approaching consciousness from a position of neuroscience, one can conclude that it must be linked to the firing of neurons, the cellular unit of the nervous system that communicates with the body through electrical patterns. From further scientific research into the neural aspects of consciousness, the most likely director of consciousness in the brain is the claustrum, a thin sheet of highly active neurons that stretches to almost all cortical areas of the brain. These neurons can be considered the source of consciousness if they meet these criteria: "(1) Be central in the connection scheme of the human brain, not too close to primary sensory or motor areas. (2) Involve several sensory areas, since consciousness integrates several sensory modalities. (3) Have activity correlated with conscious experience, even in situations where it is dissociated from direct sensory input (for instance during the perception of visual illusions) (Crick & Koch, 2003)." The claustrum fits this quite well with its quick-firing neurons and access to sensory information, making it an excellent model for AI when trying to derive consciousness.

AI Attempts at Consciousness

To better understand consciousness, scientists have been using the claustrum as a model for neural nets (see Figure I), a method in AI used to teach machines to process and experience data/information the way humans do. Neural nets have found great success in modeling the "complexity of thought, in this view, is then measured by the range of smaller abstractions you can draw on, and the number of times you can combine lower-level abstractions into higher-level abstractions — like the way we learn to distinguish dogs from birds,"(Hartnett, 2019, p.1). A tech company called SpiNNaker has programmed a computer that functions fully off of these neural nets, becoming the most accurately functioning model of consciousness. This computer simulates synapses, or spikes, in the brain and has been used to study aspects of the basal ganglia to better understand Parkinson's Disease. This is the closest humans have come to calling AI 'conscious' because of its functioning similar to the human brain, implying that we can only gauge consciousness if we have a human comparative standard. The need to understand what makes a human conscious to determine if something else is conscious is interesting in the face of innovations such as Neuralink and other brain-computer interface technologies that aim to merge the human brain and computers. Considering our very limited knowledge of consciousness and its rather controversial definition, a human-AI hybrid is enough to throw the whole concept into a loop. By no means can we say we have created conscious AI, but when merging it with the human brain, would this incorporate the AI into the human conscious experience, and if it is, does the synergy of a conscious brain with a nonconscious implant make the human more conscious? The relationship between these questions is similar to that of a square and rectangle: AI can be merged into consciousness without making the human more conscious, but the human cannot become more conscious unless the AI is enveloped into consciousness.

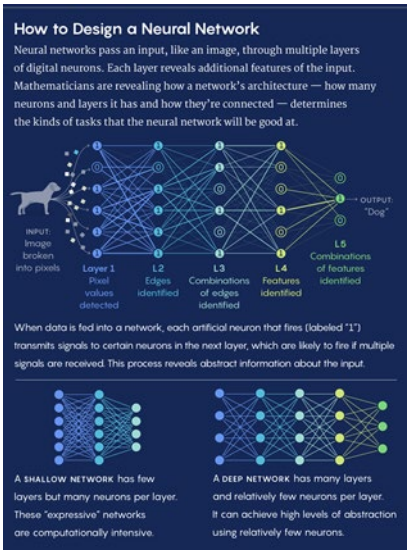

Figure 1. (Hartnett, 2019, p.1)

HEC and Information Integration Theory

Some functionalists have come up with the Hypothesis of Extended Consciousness (HEC), which suggests that our consciousness extends beyond the limits of the skull and body. This is not to be confused with a dualist perspective, as that dictates that the mind itself is outside skeletal limits, HEC acknowledges that our consciousness must extend to our environment. For example, if someone flinches when their dentist begins drilling their tooth, later the sound of the drill may cause the same reaction without the drill being physically engaged or

even near them. It is somewhat akin to classical or operant conditioning where external stimuli become part of our mental and physical reactions. Using HEC, a functionalist can say that a neural implant falls under the category of extended cognition/consciousness as it encompasses outside signals and improves awareness of inanimate events (BlueTooth wireless prosthetics, etc). But we are faced with another issue, where do we draw the line as to what in our environment is part of our consciousness? A solution to this would be Information Integration Theory (IIT), " a cognitive theory that is primarily concerned with how an individual integrates information from two or more stimuli to derive a quantitative value… IIT is developed around four interlocking psychological concepts: stimulus integration, stimulus valuation, cognitive algebra, and functional measurement (Anderson & Foster, 2014, p. 6)." Consciousness has a level of randomness that makes it so difficult for humans to replicate in AI, previously unobservable processes making up our complex functions are now quantitated by IIT theorems (shown in Figure II). We can put reasonable limits on our cognition, rather than encompassing everything within our surroundings as an unbounded function, consciousness can have a limited domain to what is relevant in our surroundings. The aspect of AI implants does not necessarily change that, but it does add more components to our relevant perception, endorsing the HEC and idea of extended cognition because something entirely inorganic becomes a part of the key element to human consciousness.

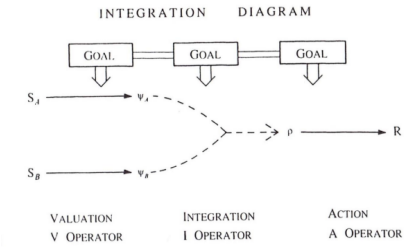

Figure 1 IIT design

	Factor A			
Factor B	$\psi_{A1}+\psi_{B1}$	$\psi_{A2}+\psi_{B1}$	···	$\psi_{An}+\psi_{B1}$
	$\psi_{A1}+\psi_{B2}$	$\psi_{A2}+\psi_{B2}$	···	$\psi_{An}+\psi_{B2}$
	$\psi_{A1}+\psi_{B3}$	$\psi_{A2}+\psi_{B3}$	···	$\psi_{An}+\psi_{B3}$
	$\psi_{A1}+\psi_{B4}$	$\psi_{A2}+\psi_{B4}$	···	$\psi_{An}+\psi_{B4}$

Figure 2 Example Factorial Design Using Additive Cognitive Algebra Model

Figure 2. (Foster, 2014, p. 64)

References

Waltz, E. (2020, January 20). *How do neural implants work?* IEEE Spectrum. https://spectrum.ieee.org/what-is-neural-implant-neuromodulation-brain-implants-electroceuticals-neuralink-definition-examples

Tononi, G. (2015, January 22). *Integrated information theory.* Scholarpedia. http://www.scholarpedia.org/article/Integrated_information_theory

Maimon, A., & Hemmo, M. (2021). *Does Neuroplasticity Support the Hypothesis of Multiple Realizability*? http://philsci-archive.pitt.edu/19174/1/Does-Neuroplasticity-Support-the-Hypothesis-of-MR-2021.pdf.

Foster, Christopher C., "The Application of Information Integration Theory to Standard Setting: Setting Cut Scores Using Cognitive Theory" (2014). Doctoral Dissertations. 39. https://doi.org/10.7275/5474959.0 https://scholarworks.umass.edu/dissertations_2/39

Extended cognition and functionalism | Mark Sprevak. (n.d.). https://marksprevak.com/publications/extended-cognition-and-functionalism-2009/

Hartnett, K., & Quanta Magazine moderates comments to facilitate an informed, substantive. (2020, May 14). *Foundations built for a general theory of Neural Networks.* Quanta Magazine. Retrieved December 23, 2022, from https://www.quantamagazine.org/foundations-built-for-a-general-theory-of-neural-networks-20190131/