Matthew Kolisnyk – 1003800050

Extended Reflection Paper – Machine Consciousness: More than just philosophically interesting.

Prepared for Dr. Christine Burton – PSY371

Over the last 70-years, Artificial Intelligence (AI) has continued to develop, as has the philosophical debate about whether it can achieve consciousness. Some proponents view this possibility as science fiction, while others believe it is only a matter of time. Whether conscious AI is just a philosophically interesting question hinges on the feasibility of consciousness in machines and what states and contents of consciousness will be available to them. If consciousness is not possible and/or the conscious contents are trivial, then this discussion is nothing more than an interesting philosophical exercise. I argue against complete dismissal of consciousness in machines by refuting typical a priori arguments against it but indicate points where skepticism appears rational. I also mention experimental evidence or possible experiments which could illustrate machine consciousness. While not clear cut, the mixed evidence points toward further investigation of machine consciousness and not its delegation to a mere philosophically interesting question.

Many different criteria have been put forward to define what is necessary to have consciousness in machines. I will split the criteria into three categories. They are phenomenal, access, and human-like consciousness. According to Block (1995), access consciousness is the ability of a system to represent, manipulate, and control information. This is separate, but interacts with, phenomenal consciousness, which is the subjective experience of information in a system. Human-like consciousness is a term I use to represent the belief that some human trait is necessary to produce consciousness. For example, this trait can be metacognition, emotion, volition, empathy, and/or some combination of these traits. Under this view, some or all of these traits are necessary for a machine for it to be conscious.

Before attempting to determine whether machine consciousness can occur, one can doubt how one would know it if one saw it. Since appearances of phenomenal consciousness in

machines could be wholly explained by an unconscious machine learning process why should one engage in this research question (Dehaene, Lau & Kouider, 2017)? This argument could be true; however, the same argument could be used against consciousness in humans. In fact, this type of argument is very similar to how behaviourism attempted to discredit cognition. Compare the similarity of Morgan's Canon to the counterargument against machine consciousness, "In no case may we interpret an action as the outcome of the exercise of a higher psychical faculty, if it can be interpreted as the outcome of the exercise of one which stands lower in the psychological scale." (Morgan, 1894). It would follow that consciousness, a higher psychology faculty, can be wholly defined in terms of unconscious processing, a lower psychology faculty. However, if one took this canon seriously, then one could never prove the existence of consciousness in humans, let alone in machines. One would only be justified in saying that other humans appear conscious while actually being composed of lower unconscious states (or some form of learning). This is, of course, false, and humans can be sure that some processing is conscious, and this logic can be safely extended to the possibility of machine consciousness.

Another worry is that phenomenal consciousness can only occur in biological systems. This appears unlikely given the trivial differences between biological and machine information processing. In fact, some theories of consciousness point toward the complexity of information integration, regardless of the underlying substrate, being the cause of consciousness (Tononi, 2004). Neurons receive input and fire in an all or none fashion, and a collection of these neurons firing in a specific way constitute our every experience. Similarly, AI is grounded in computers which are composed of many transistors, which take on either a 0 or 1, in response to an input. Furthermore, AI, through neural networks, activate output units according to interactions between input units and hidden units (Rumelhart & McClelland, 1986). Considering

computational similarities, the burden of proof lies on consciousness researchers to illustrate what properties of cells, chemicals, or other biological structures facilitate consciousness. Despite this, consciousness research has mainly been focused on neural correlates of consciousness (NCC) and not special properties of these correlates (Koch, Massimini, Boly & Tononi, 2016). Therefore, there is no a priori evidence to dismiss consciousness in AI due to it not being a biological system.

A different approach to biological systems hypothesis is to look towards why consciousness may have emerged in humans. Given that humans are the product of evolution one can wonder if there is a link between consciousness and evolution. It appears beyond coincidence that our contents of consciousness tend toward various evolutionarily ends. Sex is pleasurable, ostracism upsetting, rotten food has a bad smell while good food has a good smell, and death and disease is painful. This has led some to believe that consciousness is exclusive to living things and think that chances of it manifesting in machines is remote (Seth, 2017). For example, in the case of death being painful, machines likely do not have this conscious content due to having no preference for being turned on or off. Self-preservation is a goal in human so it could, in principle, be hard-coded in machines. For example, consider an AI which is trained to respond in a specific way to inputs. If they respond correctly, they continue to exist; else their power source is terminated for some length of time. However, without the desire for self-preservation, it appears unlikely that this content will have phenomenal quality or if self-preservation in isolation is enough to cause it. This represents a problem for machine consciousness and a blind-spot in current research.

Some argue that consciousness requires some combination of human-like traits such as emotion, metacognition, selfhood, and/or empathy (Haladjian & Montemayor, 2016). While

some human-level traits are necessary for consciousness, others appear parochial. For example, one mistake that one could make is to insist that the phenomenological content of machine consciousness must match human phenomenological contents. This is needlessly anthropomorphic as even if one grants that machines are not capable of human-type phenomenology, this says nothing about the possibility for other "non-human" states of phenomenal consciousness. For example, a python has a pit organ which allows it to detect infrared radiation emanating from prey (Fang, 2010). Despite having no human experience of this type of phenomenological experience, it would be wrong to a priori discredit its existence. However, there may be some minimum level of sensory experience (or integration) which facilitates conscious experience.

In machine consciousness, the types of conscious contents available to them may outstrip even that of humans. David Deutsch (1986) the father of quantum computing, proposed that if human-level AI interfaced with a quantum computer, it would be able to experience wave interference. Rather than merely being available to humans it seems that conscious contents are constrained by the type of sensors and perceptual processes available to the system. Further, other specific human experiences viewed as being central to consciousness such as emotions, empathy, and motivation, may not be under this interpretation. For example, one can imagine a brain lesion which completely removes all emotions, yet phenomenal consciousness still occurs due to sense modalities and thought. Hence, not all human traits are necessary for consciousness to manifest in machines and certain contents of consciousness are likely to exist beyond those available to humans.

However, there appear to be other human traits such as selfhood and metacognition which appear necessary for consciousness, regardless of the system. Selfhood is the subject to

which one prescribe our actions and thoughts. Although, I should point out, some doubt the existence of the self even in humans and therefore doubt its necessity for consciousness in general (Harris, 2014; Hood, 2012). While appearing contradictory, some believe only an experience of the self (even if false) is necessary for consciousness to emerge. Metzinger (2000) believes that this occurs through an internal self-model confusing the representation of one's body as one's actual body and prescribing one's self erroneously to that representation. In the case of AI, researchers have implemented self-models to varying degrees (Reggia, 2013). For example, Samsonovich, Kitsantas, Dabbagh and De Jong, (2008) proposed a single internal model of the self which monitors the current state of the machine by receiving inputs from a secondary perceptual unit. Similarly, Ramamurthy and Franklin, (2011) conceived their LIDA model which posits multiple selves including the proto-self (representing the current state), the minimal self (representing the self as a subject, experiencer, and agent) and extended self (representing autobiographical self, self-concept, volitional self, and narrative self). An interesting consideration is to wonder about the role of embodiment in machine consciousness and its importance for creating a reference for the self-model. One can also ask if a self-model can emerge without having an external body or environment to refer to? More research is needed to understand the role of the self in machine consciousness.

Closely related to self-image is metacognition which is the ability to know that you are thinking. This has been largely absent in current machines given the narrowness of the tasks they have to perform. The problem of common knowledge exemplifies the state of machines which struggle with very simple questions outside of their expertise (Davis & Marcus, 2015). This is a problem of cognitive flexibility and dealing with novelty and instantiating metacognition in machines may be a way of dealing with that problem. Schmill and colleagues (2008) have

recommended the addition of a Metacognitive Loop (MCL) which takes in state information of a system compares it with goals of the system and, in the case of discrepancies, outputs recommendations for the system. Observing one's state and making the appropriate changes is also a part of metacognition. This has been observed with Google's AutoML project, which outputs improved neural networks of itself by producing "child" versions that performed better on a subsequent image identification task (Le & Zoph, 2017). While it appears that these AI have some ability to reason about one's state in accordance with one's goal, one can doubt whether this is accompanied by a phenomenological component. It is possible that these narrow AI do not have complicated enough goals or states, or phenomenal experience of this state, in order to experience consciousness. This could be tested in general intelligence AI with more developed goals and input. Some have proposed the cognitive architecture which would be required to complete this task (see Crowder, Friess & Ma, 2011). Regardless, metacognition, in addition to selfhood, appears a reasonable human-like trait to base future classification of consciousness in machines.

In conclusion, several traditional attacks on consciousness have been called into question. Claims that machine consciousness is unobservable and thus not worthy of study fail in that one can recognize that humans are conscious despite not observing it directly. Biological systems are very similar structurally to advanced machine systems which make it difficult to see what may cause consciousness in one and not the other. However, the biological system's hypothesis makes a better case for doubting machine consciousness when considering the evolution as the origin of consciousness. By consider human-like consciousness, one can question if a machine is consciousness if it is capable of metacognition and selfhood but need not occur for any specific sensory modality, emotions, or motivation. Hence, the progress made in machine consciousness

makes it worthy of future study and moves it away from only being a philosophically interesting question.

References

Block, N. (1995). On a confusion about a function of consciousness. Behavioural and Brain Sciences, 18, 227-287.

Crowder, James & Friess, Shelli & Ncc, Ma. (2011). Metacognition and Metamemory Concepts for AI Systems. Conference Paper.

Davis, Ernest, Marcus, Gary (2015) Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence Communications of the ACM 58(9), 92-103 DOI: 10.1145/2701413

Deutsch, D. (1986). Three connections between Everett's interpretation and experiment In Roger Penrose & C. J. Isham (eds.), Quantum Concepts in Space and Time. New York;Oxford University Press. pp. 215--225

Dehaene, S., Lau, H., Kouider, S (2017) What is consciousness, and could machines have it? Science, 358(6362), 486-492 DOI: 10.1126/science.aan8871

Fang, J. (2010, March 14). Snake infrared detection unravelled. Retrieved from Nature: International Weekly Journal of Science: https://www.nature.com/news/2010/100314/full/news.2010.122.html

Haladjian, H.H. & Montemayor, C. (2016). Artificial consciousness and the consciousness-attention dissociation. Consciousness and Cognition, 45, 210–225.

Harris, S. (2014). Waking Up: A Guide to Spirituality Without Science. New York: Simon & Schuster Paperback.

Hood, B. (2012). The Self Illusion: How the Social Brain Creates Identity. New York: Oxford University Press.

Koch, C., Massimini, M., Boly, M. & Tononi, G. (2016) Neural correlates of consciousness: progress and problems. Nature Reviews Neuroscience, 17, 307–321

Le, Q. & Zoph, B. (2017) Using Machine Learning to Explore Neural Network Architecture. Retrieved from https://ai.googleblog.com/2017/05/using-machine-learning-to-explore.html

Metzinger, T. (2000). The subjectivity of subjective experience T. Neural correlates of consciousness, MIT Press.

Ramamurthy, Uma & Franklin, Stan. (2011). Self System in a Model of Cognition. International Journal of Machine Consciousness 4(2) doi:10.1142/S1793843012400185.

Reggia, J. (2013). The rise of machine consciousness: Studying consciousness with computational models. Neural Networks 44, 112-131

Rumelhart, D., & McClelland, J. (1986). Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations. Cambridge: MIT Press.

Samsonovich, Alexei, Kitsantas, Anastasia, Dabbagh, Nada & De Jong, Kenneth. (2008). Self-awareness as metacognition about own self concept. AAAI Workshop - Technical Report.

Schmill, M.D., Oates, T., Anderson, M.L., Josyula, D., Perlis, D., Wilson, S., and Fults, S., (2008) The Role of Metacognition in Robust AI Systems. AAAI Workshop - Technical Report

Seth, A. (2017). Your brain hallucinates your conscious reality [Video File] Retrieved from https://www.ted.com/talks/anil_seth_how_your_brain_hallucinates_your_conscious_reality/footnotes?fbclid=IwAR1MCudlA7YuyLQRiNZvYs2XsQ57cT2tuVhfgPyfnX5rrMDEHFzSj10XWh8&utm_campaign=tedspread&utm_medium=referral&utm_source=tedcomshare

Tononi, G. (2004) An information integration theory of consciousness. BMC Neuroscience 5(42) https://doi.org/10.1186/1471-2202-5-42