proceedings

MDPI

Proceeding Paper

Consciousness, Machines, and Ethics [†]

Jim Davies

Department of Cognitive Science, Carleton University, Ottawa, ON K1S 5B6, Canada; jim@jimdavies.org

† Presented at the Conference on Theoretical and Foundational Problems in Information Studies, IS4SI Summit 2021, online, 12–19 September 2021.

Abstract: The field of consciousness displays a striking lack of consensus on many important issues, one of which is the possibility of consciousness in machines and software. However, consciousness in machines has dramatic ethical implications, regardless of which ethical framework is used. This suggests that the best course of action would be to try not to create conscious machines until we better understand the issue. However, understanding machine consciousness without building it might not be possible.

Keywords: consciousness; artificial intelligence; ethics; artificial general intelligence; artificial superintelligence; machine ethics; morality; sentience

check for updates

Citation: Davies, J. Consciousness, Machines, and Ethics. *Proceedings* **2022**, *81*, 40. https://doi.org/10.3390/proceedings2022081040

Academic Editor: Mark Burgin

Published: 19 September 2021

1. Introduction

In this paper, I will discuss some issues with creating conscious artificial intelligences. Very few people are specifically trying to make conscious artificial intelligence. The reason for this is that consciousness is not particularly useful, as far as we know. We really do not know what consciousness is for human beings, and many prominent consciousness researchers do not believe it has any function at all: one of the most prominent theories of consciousness holds that it has little or no function [1]. Regardless, we are even more in the dark about why an artificial intelligence (AI) might need consciousness because consciousness, as far as we know, has no useful purpose for artificial intelligence. Rather than trying to make conscious machines or software, AI researchers are focused on making useful applications.

Even if, in the future, we were to find that consciousness has a function and is, perhaps, necessary for some cognitive functions in humans, it does not then follow that artificial intelligences would also need consciousness to do those same tasks. As a result, the nature of conscious artificial intelligence is heavily understudied in the artificial intelligence community.

A major issue with the study of conscious artificial intelligence is that we have no good measure of consciousness in the machine. There are several theories of consciousness, and each one of them has some predictions that we could use to determine whether or not a given machine was conscious. Unfortunately, we do not have a theory-independent measure of consciousness. This is of course a problem for evaluating these different theories of consciousness, and so it might be that we need to have a very good theory of consciousness before we can come up with a decent test of consciousness in machines.

Let us look at one popular example, the Turing test, which is commonly thought to serve as a test of consciousness in machines. We do not really know whether or not the ability to effectively communicate with language requires a machine to have subjective conscious states, and the three most promising theories of consciousness do not make any special claims about the relationship between language and consciousness.

Now I will discuss these three prominent theories of consciousness and how they relate to machines. The field of consciousness studies is incredibly controversial and rife with uncertainty. A recent survey of consciousness scholars shows that there were three contemporary theories that are very promising [2]. The most popular theory, which is called global workspace theory, was thought to be particularly promising, with about 36% of researchers reporting that it was a promising theory. The second most popular theory is higher-order theory, which had a 17% approval rating. The third most popular was integrated information theory, with 14%.

The first thing to notice about these percentages is that they are all fairly low, and there was nowhere near a consensus on what theory is correct. It is very possible that none of these theories are correct. I will briefly describe each of them.

2. Popular Theories of Consciousness

2.1. Global Workspace Theory

First we have the most popular, global workspace, and its neural counterpart, global neuronal workspace [3]. In this theory, the mind can process things unconsciously, but these processes are relatively local. In order for a thought to be widely accessible to many processes in the mind, it needs to be shared like a viral email across the cortex: global ignition. According to global workspace theory, this constitutes consciousness. There is a certain threshold that appears to be reached in processing where the entire cortex repeats the same information, and this is thought to be what consciousness consists of: an all-or-nothing effect of information being shared in a temporary memory buffer that makes the information accessible to many functions of mind.

What does this mean for artificial intelligence? Most of the scholarly work on global workspace theory does not consider artificial intelligence at all, but what there is suggests that a system might have the capacity to be conscious if it had local processing functions, but also a shared "blackboard" buffer where information can be shared globally [4]. Unfortunately, the criteria for what would or would not constitute a suitable blackboard buffer in an AI is unclear. Arguably, an undergraduate student could create an artificial intelligence system that had local and global properties for a course assignment. Perhaps there is some level of complexity that is required for this global workspace to constitute consciousness, but what this threshold is has not been defined or experimental determined. As it stands, it might be that some very simple artificial intelligences might already be conscious, according to this theory.

2.2. Higher-Order Thought Theory

The next most popular theory is the higher-order theory, the most popular version of which was created by David Rosenthal [5]. This is a subtle and complex theory, but for our purposes, we will focus on the simplest part of it. Suppose you hear a microwave beep. According to higher-order thought theory, the perception of this beat is unconscious at first. Later, what might happen is that your mind creates a higher-order thought, the contents of which states that the agent is in a state of hearing a beep. When a higher-order thought like this is created, that references the beep, making the person conscious of the beep. Being conscious of something is simply a matter of having a higher-order thought about that something.

When we turn to artificial intelligence, it is even easier to make artificial intelligence with higher-order thoughts than it is to make artificial intelligence with the global workspace. There is nothing particularly in the theory that says that a program I could create in an afternoon would not have the kinds of higher-order thoughts required to constitute consciousness.

2.3. Integrated Information Theory

The final theory I will discuss is integrated information theory [6]. One striking aspect of this theory that differentiates it from most others is that whether or not consciousness is present is dependent on the physical hardware or neurons of the agent in question. Integrated information theory says that consciousness is present when there are a number of units, each of which can take on different states that have the capacity to interact with each other in a recurrent way (this is an oversimplification; the mathematics describing this kind of recurrence are quite complex). This means that any strictly feed-forward mechanisms lack consciousness altogether.

Its relationship to artificial intelligence is particularly interesting: the software that the computer is running is utterly irrelevant. All that matters is the nature of the hardware. Just about all the computers that we use today are completely unconscious because of the nature of their hardware architectures. Creating a conscious machine, for integrated information theory, requires making a computer, the hardware of which consists of a series of physical units that interact with each other in a recurrent way. Then, it does not matter whether it is running Microsoft Word or artificial intelligence. On this theory, as long as we are using traditional von Neumann computer architecture, neither the computers nor the software that run on them will ever be conscious. Keep in mind that although this is the third most popular consciousness theory among researchers, it is still endorsed by only 14% of them.

3. The State of the Art of Machine Consciousness

Although I have reviewed only a fraction of the theories of consciousness, it should be clear that, if taking stock of the field as a whole, there is enormous disagreement on what might make a machine or the software that runs on it conscious. This is supported by an informal review, which reveals that there is no consensus among the top consciousness scholars on this issue [7].

What should we take from this? Although one feels particularly confident about whether or not computers have the capacity, or could someday have the capacity, for consciousness, the actual state of the consciousness field should give one pause. If we respect the uncertainty that we see among experts in the field, the rational way to think about the situation is that we are very, very much in the dark about whether computers could be conscious, and if they could, how that might be achieved. Put more bluntly, it is *irrational* to put a high certainty on one's beliefs about this issue. Doing so would require one to be justified in believing that one's understanding of the issues is significantly greater than that of the field as a whole.

Depending on what (perhaps as-of-yet hypothetical) theory turns out to be correct, computers might be conscious someday, computers will never be conscious, or we already have some that are. Although few researchers are *deliberately* trying to make conscious machines, it is possible that consciousness might arise as a side-effect of attempting to implement other functions, such as memory, attention, and reward. We just do not know.

This puts us in an unfortunate position: we do not know how to make conscious machines, and without a theory-independent test of the existence of consciousness in machines, we will not know if and when we do create one, and at the same time the existence of conscious machines would have dramatic ethical consequences.

4. Ethics and Consciousness

The field of ethics is at least as controversial as that of consciousness. In ethics, at least, many views can be decently categorized as belonging to one of three major kinds.

The first is known as deontology. If you talk to most common people about their ethical views, they will probably communicate those views in terms of rules. For example, you are not allowed to kill, and you should take care of your children. Many ethicists are also deontologists.

Another popular theory is utilitarianism, which avoids thinking in terms of rules, and instead endorses the maximization of well-being.

The last major categorisation of ethical thought is virtue ethics, a theory that suggests that you should consider each situation you encounter in terms of a small set of ethical virtues, such as humility, honesty, and justice.

Luckily for us, the relationship of consciousness to ethics is somewhat orthogonal to these debates. In short, many people from all ethical camps believe that consciousness is an important part of determining what things in this universe are worthy of moral consideration in which are not. Put baldly, the fact that a human being can enjoy pleasures and suffer pains is relevant to the fact that a person can be wrong and treated poorly. A rock, which is not conscious, cannot be wronged, and, for many, the fact that the rock is completely unconscious is a relevant and important reason why.

It is not just any kind of consciousness that is relevant to ethics. Specifically, the kind of consciousness that matters is *valanced* consciousness; that is, it is conscious states can be pleasant or unpleasant. It might be that a being could be conscious of something, such as a perception of a microwave beeping, but not be conscious of any good or bad along with it. Completely neutral conscious thoughts are, under most ethical structures, ethically irrelevant. In ethics, we call beings that can feel good or bad conscience feelings *sentient*, and beings that deserve moral consideration *moral patients*. In most ethical frameworks there is considerable overlap between these sets.

Sentient Computers or Software as Ethical Patients

This means that the sentience of computers and software is of very high ethical importance. Specifically, if we were to create a sentient computer, we would have ethical obligations to treat it well. This is much more complicated than it might first seem.

How would we know what events would increase or decrease the machine's welfare? Consider that we create computers and software to do work we cannot or do not want to do. To the extent that those machines are sentient suggests we would be required to give ethical consideration to how much the execution of those tasks affect the machines' welfare. Unless we can be sure that machines would actually *enjoy* the work we have them do, it would be unethical to make them sentient. Perhaps we will someday be able to program them so that they really enjoy exactly the jobs we want them to do.

Another problem is the computer analogue of murder: if one had sentient software running on one's laptop, would we ever be justified in turning off the computer? It becomes worse: we can look at computers at three levels: the computer itself, the software we can run on it, and then every *instance* of running that software. That is, when you ran Microsoft Word this morning, it is a different instance of running it than when you ran the same program two years ago. To which level do we have ethical commitments? Might we need to keep each *instance* of sentient software running indefinitely, even if it is inefficient and producing very low welfare?

All of this suggests that we probably should not create sentient machines if we can help it. (Conscious machines might be ethically built if they are not sentient).

Now I will turn that on its head. If machines can have conscious welfare, they also might be able to produce it more efficiently than biological beings. That is, for a given number of resources, one might be able to produce more happiness or pleasure in an artificial system than any living creature. Suppose, for example, a future technology would allow us to create a small computer that could be happier than a euphoric human being, but only required a cell phone's amount of energy out of a wall socket. According to utilitarianism, this might lead to the conclusion that our eventual best course of action would be to create as much artificial welfare as we can, turning as much matter in the universe into machines that efficiently produce welfare, perhaps 10,000 times more efficiently than it can be generated in any living creature—a theoretical substance philosophers call "hedonium" [8].

This suggests that we should not make conscious machines until we understand them well enough to create them deliberately for the purpose of generating welfare. Unfortunately, it also might prove to be difficult to understand consciousness without modeling it on computers, as modeling is a valuable way to explore and test theories in psychology. That is, we might not ever know how to create a conscious machine without actually trying to build them, for the same reasons it would have been difficult to learn how to build large bridges without experimentation. In any case, research on conscious machines has a strong ethical component fraught with uncertainty.

Data Availability Statement: Not applicable.

References

1. Rosenthal, D.M. Consciousness and its function. *Neuropsychologia* **2008**, *46*, 829–840. [CrossRef] [PubMed]
2. Francken, J.; Beerendonk, L.; Molenaar, D.; Fahrenfort, J.; Kiverstein, J.; Seth, A.; van Gaal, S. An academic survey on theoretical foundations, common assumptions and the current state of the field of consciousness science. *PsyArxiv* **2021**. Available online: https://psyarxiv.com/8mbsk/ (accessed on 11 March 2022).
3. Mashour, G.A.; Roelfsema, P.; Changeux, J.P.; Dehaene, S. Conscious processing and the global neuronal workspace hypothesis. *Neuron* **2020**, *105*, 776–798. [CrossRef] [PubMed]
4. Craig, I.D. Blackboard systems. *Artif. Intell. Rev.* **1988**, *2*, 103–118. [CrossRef]
5. Lau, H.; Rosenthal, D. Empirical support for higher-order theories of conscious awareness. *Trends Cogn. Sci.* **2011**, *15*, 365–373. [CrossRef] [PubMed]
6. Oizumi, M.; Albantakis, L.; Tononi, G. From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS Comput. Biol.* **2014**, *10*, e1003588. [CrossRef] [PubMed]
7. Blackmore, S.J. *Conversations on Consciousness: What the Best Minds Think about the Brain, Free Will, and What It Means to Be Human*; Oxford University Press: Oxford, UK, 2006.
8. Schwitzgebel, E.; Garza, M. A defense of the rights of artificial intelligences. *Midwest Stud. Philos.* **2015**, *39*, 98–119. [CrossRef]