Why Artificial Consciousness Matters*

Matthew Crosby[0000−0001−7606−7533]

Imperial College London
Leverhulme Centre for the Future of Intelligence
m.crosby@imperial.ac.uk

Abstract. In this paper I outline the case for artificial consciousness research which, while rising in popularity, remains less impactful than work in neighbouring fields. I argue that the ethical concerns around creating artificial consciousness are important, even with respect to near-future incremental advances towards general intelligence. I also give examples of how AI consciousness in particular provides unique tools and methods for testing and thinking about theories of consciousness from both metaphysical and practical standpoints.

Keywords: Artificial Consciousness · Phenomenal Consciousness · Ethics

1 Introduction

While consciousness research has progressed rapidly in recent years, scientific consensus remains out of reach. Meanwhile, AI has made significant advancements, reaching superhuman performance on a wide range of tasks. Humans are no longer the best quiz-show contestants [8], the best Go players [18], or even, in some aspects, the best doctors [7]. This rapid expansion, and Bostrom's influential discussion of mind crimes [2], has led some artificial consciousness research to be framed partially [1], or completely, in terms of superintelligence [15, 21].

However, even with recent advances, no current state-of-the art AI approaches can compete with simple animals when it comes to adapting to unexpected changes to their environment or any of a wide range of behaviours commonly associated with consciousness. Furthermore, when we attribute consciousness to animals, it is generally accepted that human-like intelligence is not necessary. The first conscious AI systems are not likely to have surpassed the general intelligence of humans, or even of many animals usually considered conscious [12]. Understanding such systems of minimal artificial consciousness presents important ethical questions, and also opportunities to improve both our understanding of AI and of our own consciousness.

Whilst the amount of work on artificial consciousness has been growing, the area has yet to be as impactful as other fields in either consciousness or AI [13]. This paper presents the case for the importance and timeliness of research into artificial consciousness without superintelligence. It is ethically relevant, it can teach us about consciousness, and it can help us progress towards general intelligence in AI.

* Work supported by the Leverhulme Centre for the Future of Intelligence.

2 Phenomenal Consciousness

For the purposes of this paper I will restrict discussion to phenomenal consciousness. There is no agreed upon definition. For our purposes I refer to Schwitzgebel's extended definition by example [16]. Positive examples include your current visual experience of this text, the auditory experience of surrounding sounds, any imagery you choose to conjure into the mind, and the felt sense of emotions. Negative examples include your current visual experience of the other side of the world, currently unrecalled knowledge (such as that Paris is the capital of France), and the flow of blood in the capillaries of your left knee. Phenomenal consciousness is then defined as the simplest thing common to all the positive examples and missing from all negative ones. While there are technically an infinite number of concepts (and combinations of concepts) that fit this definition, the intended interpretation is assumed to be clear enough.

There are many proposed tests for machine consciousness, most of which ignore the phenomenological component, instead focusing on more measurable related aspects [13]. However, ignoring the phenomenological component amounts to ignoring the reason that consciousness is interesting in the first place, and the thing that sets it apart from general measures of intelligence. The problem of how phsyical systems can give rise to phenomenal experiences at all, the 'hard problem' of consciousness, must be addressed, even if only to be explained away, by any complete theory of artificial consciousness [5]. For the purposes of this paper I will assume consciousness means phenomenal consciousness and maintain the commonly held assumptions that consciousness supervenes on purely physical processes and there are many physical substrates that could realise it.

3 The case for artificial consciousness

This section is split into two parts. The first part concerns the ethical considerations around introducing potentially conscious entities to the world. I argue that discussion should be focused on near-term possibilities of minimal consciousness. The second part gives examples of the unique methodologies and insights about consciousness afforded by research into artificial intelligence.

3.1 The ethical case

There is no definitive answer to which animals are conscious. At one end of the scale, panpsychists ascribe consciousness to all animals (and beyond) [4]. At the other extreme some higher-order thought theorists deny consciousness to most non-human animals, though even here there is large disagreement [14]. Most views sit somewhere in the middle, claiming human-like and more intelligent animals are conscious whilst other 'lesser' animals are not. For example, Peter Godfrey-Smith suggests mammals, birds, squids and octopods as candidates whilst other mollusks, jellyfish, and many fish are not [9]. Categorising animals by consciousness is made even harder under the common view that consciousness

exists on a gradual scale, whereby ascriptions on a per-animal basis are more involved than a simple 'yes' or 'no' answer.

While there is wide disagreement in ascriptions of consciousness, there is one thing common to all the above; superintelligence is not required. The human case by itself is enough to prove this. The majority of theories go much further, ascribing consciousness to entities with rudimentary - or even in some cases, no - intelligence. Extrapolating to artificial consciousness, it may be possible to recreate consciousness in minimally intelligent artificial systems.

Furthermore, it is likely that we will create minimally intelligent conscious systems long before we create entities that have, or exceed, human-like intelligence. As AI progress incrementally towards its goal, the question of AI consciousness will arise for, at least in the biological sense, minimally intelligent systems long before it does for superintelligent systems. Measured on aspects of intelligence common in biologocial systems, the first ever conscious artificial system will likely be closer to a mouse than a superhuman.

Thomnas Metzinger has put forward a thought-experiment in which he suggests that the first conscious AI systems would find themselves in a situation similar to that of human infants with cognitive and emotional deficits [12]. The idea is that if any AI systems are ever conscious, the first set of conscious entities will be the least capable that we can create. As with any engineering project, the first few attempts will be less streamlined and capable than future versions. In comparison to our consciousness, they will be both cognitively and emotionally deficient, which leads to obvious concerns about suffering.

I take the general case for the importance of ethical considerations towards any conscious entities with the capacity for capacity for suffering to be uncontroversial [19]. Should it be possible to create entities that suffer, then they will naturally have moral patiency. The real question is whether or not we will ever have even minimal artificial consciousness and, even if yes, whether it will be of a form with the capacity to suffer. Sotala et al specify three general rules for a problem to be worth considering (paraphrased below) [21]. They apply these to the superintelligence case, but the rules by themselves are also applicable here:

1. The outcome must cause enough harm to merit attention.
2. The outcome must have reasonable probability of being realised.
3. There must be some way to reduce either 1. or 2.

If it can be shown that just one of these conditions is violated then a problem becomes much less pressing. However, for the case of minimally conscious AI, I will argue that all of these hold.

The scale at which AI systems can be created or destroyed suggests the problem is 'bad enough'. Should it be possible to create a conscious artificial entity, then its artificial nature will allow for easy copying and replicating. Many modern AI algorithms involve creating multiple instantiations and then testing them over a wide range of domains. This practice means that any created AI consciousness would probably exist many times over, multiplying the ethical significance of the event.

Whether or not even minimal artificial consciousness has a reasonable probability of being realised is still open to debate. However, there are many reasons to believe this will happen at some point and AI research is making incremental progress in this direction. Given the ubiquitous nature of biological consciousness and the progress of cognitively inspired AI [10, 11], it is premature to rule out the possibility of artificial consciousness, especially accidentally or unknowingly.

Finally, there must be some way to reduce the probability of causing harm or the amount of harm caused. As our understanding of consciousness grows, our ability to influence its creation also grows. If we do develop a better understanding of which systems are conscious, this could be used to, in the extreme case, avoid creating such systems in the first place.

Artificial consciousness research reasonably meets all three of the criteria, but, more importantly, the criteria are useful for shaping the most important areas of research. The most ethically abhorrent scenario is that artificial consciousness research lags behind AI progress and we only find out too late that we have been creating conscious, suffering AI systems. This makes the topic not only ethically relevant, but urgent. Our understanding of consciousness must keep pace with, and be applied to, progress in AI.

3.2 Learning about consciousness from AI

Beyond the ethical considerations, AI is uniquely placed to teach us about consciousness. It can be a testbed for theories, and also provide insights into longstanding metaphysical problems.

The Roomba Test One use from considering theories of consciousness applicable to AI systems is that they can be evaluated through minimal instantiations. Any substrate independent theory of consciousness has to answer to its simplest possible implementation in AI, and often that implementation will provide insights into what is working and what is missing from the theory. I have called this the Roomba Test because this well-known device features in public discourse on machine consciousness [20] due to meeting some rudimentary properties related to consciousness; autonomy and representations of the world. At its simplest this test can be just a thought experiment where it serves as a sanity check on theories of consciousness. However, actual implementations provide much more feedback and can then be further compared against preferred measures of consciousness.

One response to Roomba examples is to bite the bullet, such as IIT does [22] when faced with similar questions about systems that, unlike the Roomba, would have non-zero consciousness (i.e. $\Phi > 0$) [17]. The theory predicts that a small set of interconnected logic gates is conscious, and, in the face of such a system, sticks to this claim. Another response is to use such examples to refine the theory and move towards a better understanding as is the dialectical method in the presentation of the radical plasticity thesis [6]. If there exists no simplest instantiation of an artificial system then this stands as a critique of the generality of the theory. Either way, AI provides an important tool for evaluating the claims of theories of consciousness.

The Veridical Dualist Most of modern day AI research uses simulated environments for testing due to their many advantages over real-world, or static, data sets. Simulated environments can be constrained to test for exactly the problem that is being searched, noise can be controlled, they can be simplified as necessary, mistakes don't have real-world consequences, and they are cheap to copy and run. Algorithms can be swapped in and out and are usually run in a separate thread (and often using completely different programming languages) to the simulated environments themselves. With all these advantages, and the recent proliferation and improvement of such environments it is likely that more and more systems will be tested in this way in the future.

The veridical dualist thought experiment, (originally from Chalmers [3]), asks you to imagine you are an artificial entity, existing in a rich simulated environment. Furthermore, you happen to be conscious and even capable of asking (and solving) questions about your consciousness. You have learnt much about your simulated world by exploring it and probing it, paying close attention to how it reacts and moves. All this information becomes integrated into your experience of the simulated environment. However, it occurs to you that the thoughts and experiences you have don't seem to originate from within this world. They seem (comparitively) ethereal. They seem to come from somewhere else, perhaps they are some kind of secondary metaphysical type to everything else in your (unbeknownst to you) simulated environment.

A simulated entity might naturally become an ontological dualist. No matter how hard the artificial entity looks for the underlying substrate forming its experiences inside its world, it will never be able to find them. Its equivalent of a brain is running in a computer in our world and, presumably, giving rise to conscious experiences that, from its perspectives, exist in the only world it knows, the simulated one. It seems rational for it to become a dualist and, from the perspective of its simulated world, correct to do so.

What this thought experiment, and the host of others that can be created by considering the unique properties of simulated environments, tell us about our own situation is not yet clear. This certainly does not mean we should all be dualists, but, as Chalmers put it, perhaps "dualism isn't quite so outlandish and conceptually problematic as tends to be supposed." The important point here is that AI consciousness provides some unique perspectives for advancing our understanding and will continue to be fruitful on this front as we come closer to realising complex simulated worlds and able to implement more complicated theories of consciousness.

4 Conclusion

In summary, there are many reasons to expect that AI will play an important role in our future understanding of consciousness. Furthermore, understanding artificial consciousness is an ethically relevant problem on a scale only limited by our increasing computational resources. It will be important in the future that advances in consciousness science are translated and incorporated into our understanding of artificial consciousness. It will also be important that, in the

other direction, advances in AI are translated back to consciousness science, not least, to ensure that if we ever do create artificial consciousness, we have at least some idea of what we are doing.

References

1. Arrabales, R., Ledezma, A., Sanchis, A.: Consscale: A pragmatic scale for measuring the level of consciousness in artificial agents. Journal of Consciousness Studies **17**(3-4), 131–164 (2010)
2. Bostrom, N.: Superintelligence: paths, dangers, strategies. Oxford University Press Oxford (2014)
3. Chalmers, D.: How cartesian dualism might have been true. http://consc.net/notes/dualism.html, accessed: 2018-11-01
4. Chalmers, D.: Panpsychism and panprotopsychism. Consciousness in the physical world: Perspectives on Russellian monism **246** (2015)
5. Chalmers, D.J.: Facing up to the problem of consciousness. Journal of consciousness studies **2**(3), 200–219 (1995)
6. Cleeremans, A.: Consciousness: the radical plasticity thesis. Progress in brain research **168**, 19–33 (2007)
7. Fauw, D., et al.: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature medicine **24**(9), 1342 (2018)
8. Ferrucci, et al.: Building watson: An overview of the deepqa project. AI magazine **31**(3), 59–79 (2010)
9. Godfrey-Smith, P.: Other minds: The octopus, the sea, and the deep origins of consciousness. Farrar, Straus and Giroux (2016)
10. Hassabis, D., Kumaran, D., Summerfield, C., Botvinick, M.: Neuroscience-inspired artificial intelligence. Neuron **95**(2), 245–258 (2017)
11. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. Behavioral and Brain Sciences **40** (2017)
12. Metzinger, T.: Two principles for robot ethics. Robotik und Gesetzgebung pp. 247–286 (2013)
13. Raoult, A., Yampolskiy, R.: Reviewing tests for machine consciousness (2015)
14. Rosenthal, D.M.: Varieties of higher-order theory **56**, 17–44 (2004)
15. Schneider, S.: Superintelligent AI and the postbiological cosmos approach. What is life (2017)
16. Schwitzgebel, E.: Phenomenal consciousness, defined and defended as innocently as i can manage. Journal of Consciousness Studies **23**(11-12), 224–235 (2016)
17. Shanahan, M.: Ascribing consciousness to artificial intelligence. arXiv preprint arXiv:1504.05696 (2015)
18. Silver, D., et al.: Mastering the game of go without human knowledge. Nature **550**(7676), 354 (2017)
19. Singer, P.: Practical ethics. Cambridge university press (2011)
20. Singler, B., Smith, E.S.J., Ramsay, C., Uren, J.: Pain in the machine: short film
21. Sotala, K., Gloor, L.: Superintelligence as a cause or cure for risks of astronomical suffering. Informatica **41**(4) (2017)
22. Tononi, G., Boly, M., Massimini, M., Koch, C.: Integrated information theory: from consciousness to its physical substrate. Nature Reviews Neuroscience **17**, 450–461 (July 2016)