JOURNAL OF SOCIAL COMPUTING
ISSN 2688-5255 02/07 pp270−274
Volume 4, Number 4, December 2023
DOI: 10.23919/JSC.2023.0023

On the Existence of Robot Zombies and our Ethical Obligations to AI Systems

Luke R. Hansen*

Abstract: As artificial intelligence algorithms improve, we will interact with programs that seem increasingly human. We may never know if these algorithms are sentient, yet this quality is crucial to ethical considerations regarding their moral status. We will likely have to make important decisions without a full understanding of the relevant issues and facts. Given this ignorance, we ought to take seriously the prospect that some systems are sentient. It would be a moral catastrophe if we were to treat them as if they were not sentient, but, in reality they are.

Key words: artificial intelligence; sentience; ethics

1 Robot Zombie

A philosophical zombie is a hypothetical creature which walks like us, talks like us, and cries when it stubs its toe[1]. These quasi-humans act exactly like normal people but, crucially, they are devoid of sentience. They cannot experience the sweetness of a cherry or the blueness of the sky. They merely act like a human who has these experiences. Assessing the sentience of complex systems is difficult. Nagel[2] colorfully described some of these difficulties in his essay, "What is it like to be a bat?". We surmise it is "like" something to be a bat. We think we can resonate with certain aspects of a bat's life, since we, too, are mammals. Yet the experience of "bat-ness" will never be available to us, because we are not bats. Sentience is subjective and private; there are no stable grounds to compare experience across species, or even across individual human beings. In this paper, the term "sentient" will refer to an entity with the ability to have subjective experiences. This means more than objectively registering facts about the state of the external world. It means experiencing an interior world.

Philosophers have used the idea of the philosophical zombie to ask questions about the nature of sentience[3]. For example: Why do we assume other people are sentient, and not philosophical zombies? Can we know that other people are sentient? How do we decide if something that seems to be sentient is, in fact, sentient?

Our perplexity is compounded if we try to apply the word and concept of sentience to artificial intelligence systems. As artificial intelligence algorithms improve, they will seem increasingly human. Already, large language models can respond to queries in ways that seem compellingly human[4]. The questions posed by philosophical zombie thought experiments will demand to be reckoned with: How will we know whether an artificial intelligence has a degree of sentience, or whether it is just a robot zombie? In this paper, I will argue for epistemological humility: It is unlikely that we will ever be able to know for sure if our machines are sentient.

2 Theory of Sentience

Many theories have been proposed about how the brain, which is an object, relates to sentience. For the purpose of this paper, I will divide the theories into two camps: substrate independent theories of sentience and substrate dependent theories of sentience[5].

Substrate independent theories posit that a physical

• Luke R. Hansen is with the Computer Science Department, Stanford University, Stanford, CA 94305, USA. E-mail: lrhansen@stanford.edu.
* To whom correspondence should be addressed.

Manuscript received: 2023-07-01; revised: 2023-11-30; accepted: 2023-12-02

system with a particular functional organization is necessary and sufficient to effect sentience. Chalmers[6], a proponent of this view, defined functional organization as "the abstract pattern of causal interaction between the components of a system". It does not matter whether the casual relationships are expressed by neurons, transistors, or Legos—as long as the system has the right functional organization, it will be sentient. One of the most popular substrate independent theories is called information integration theory[7]. It proposes that sentience is the result of sufficiently sophisticated processing and integration of information systems.

Substrate dependent theories assert that some specific physical property of the brain is necessary for sentience. In the same way, you need an object with mass to create gravity, and some physical aspect of our brain's wetware is needed for sentience. An example of such a theory is defended by physicist Penrose[8]. He believed that quantum effects occurring in microtubules (which are small neural substructures) are necessary for sentience[8].

Are substrate dependent and substrate independent theories of sentience mutually exclusive? Let us suppose they are not. In other words, sentience is both substrate dependent and substrate independent. This would imply that sentience depends on a particular substrate doing a particular type of information processing. However, this formulation is inconsistent with the definition of substrate independence. If sentience is at all dependent on the specific physical substrate (other than that substrate's ability to express an abstract pattern of causal interaction), then it is substrate dependent. Substrate independence proposes sentience is purely an emergent phenomenon of sophisticated information processing (and can be realized on any substrate that can express causal interactions). This pattern of information processing is necessary and sufficient for sentience. For substrate dependent theories, sophisticated information processing might be necessary for sentience, but it is not sufficient. Some specific physical quality of the substrate is also essential.

Now we can re-consider one of the questions posed by the philosophical zombie thought experiments: Why do we assume that other people have feelings and experiences—that they are not philosophical zombies? Certainly it would be impractical, indeed it would be practically psychopathic to assume that other people are philosophical zombies. But this leap does not have to be justified only by pragmatic considerations: Any conceivable theory of sentience that could explain one's own sentience would also explain everybody else's, excluding an assumption of solipsism. It does not matter if sentience is substrate dependent or substrate independent, because we all have similar wetware and similar information processing systems. A piece of granite does not share our wetware nor our information processing systems, so we surmise that it does not have any form of sentience.

3 Sentience in Artificial Intelligence

Now consider artificial intelligence. As artificial intelligence functions are, and will likely continue to be, realized in silicon, we should not project sentience onto them with such confidence. They would only be sentient if sentience is substrate independent—if sentience emerges from a pattern of casual interactions, then they will be sentient as long as they are expressing those relationships.

It is extremely likely we will find ourselves interacting with machines that seem sentient, without having dispositively determined whether they are sentient, that is, whether they have internal experiences. In the last five years, there has been immense progress in making artificial intelligence seem more human. For example, the output of GPT-2 (released in 2018) tends to be syntactically plausible, but it clearly does not have a robust semantic representation of the text it manipulates[9]. Its successor (GPT-3, released in 2020) is significantly more competent, both in terms of its performance on benchmarks and qualitative assessments[10]. The trajectory of the progress has continued: GPT-4 significantly outperforms GPT-3 on almost all benchmarks, and, arguably, passes the Turing test[11, 12]. Whether through scaling existing techniques or developing new algorithmic approaches, it is an extremely safe bet that this progress will continue. Most artificial intelligence experts expect human-level artificial intelligence will be achieved within a few decades[13]. Computers are getting better at acting sentient extremely fast.

In contrast to the rapid progression of artificial intelligence systems that seem sentient, the progress in reaching consensus on the fundamental nature of sentience has been glacial. There are some profound challenges to determining whether or not sentience is a substrate dependent phenomenon. Teasing out substrate dependence from substrate independence would require us to answer the question of "Why are some physical systems sentient, and others are not?" Chalmers[14] described why answering this question is so difficult (calling it the hard problem of consciousness) in his seminal paper, "Facing up to the problem of consciousness". He suggested that our current approach to scientific understanding might not be sufficient to explain sentience. Some philosophers take an even stronger position: that we will never be able to understand sentience, that it is out of reach for our limited minds[15]. In contrast, other philosophers deny that the hard problem of consciousness even exists[16]. There is no consensus on how to answer questions about the fundamental nature of sentience, or even what the right questions are.

4 Precuationary Principle

Unfortunately, whether or not non-human systems are sentient is crucial to deliberations about the moral status of these systems. For the purposes of this essay, to have a moral status means that the way we treat an entity can be described using moral terms (in other words, there are "right" and "wrong" ways of treating that entity). If an entity is not sentient, and has no potential to be sentient, then it has no moral status in itself whatsoever. This position is known as "sentientism"[17]. Of course saying that an entity has moral status does not entail that our moral obligations to that entity are equivalent to the obligations we have to other humans. Most people believe we have some moral obligations to dogs, but those obligations are quite different from the ones we avail to other humans. Furthermore, saying an entity has moral status does not entail that it has moral qualities such as free-will and moral accountability. Again, consider animals and young humans: We recognize them as having moral status, but we do not hold them morally accountable for their actions. A thorough discussion of the relationship between sentience and an entity's moral status is out of scope for this essay. For our purposes, we will merely assume that if an entity is sentient, then it has some moral status.

If sentience is substrate dependent, then we can be sure that our machines are not sentient (provided silicon does not have the specific physical property needed for sentience). However, it is unlikely that we will be able to make this conclusion by the time we are interacting with computers that seem sentient. So then, given the epistemological barriers to assessing the sentience of systems which behave as if they were sentient, how do we navigate ethical questions regarding machines that could conceivably have sentience? The very presence of such machines will force us to make decisions about how we treat them. If we avoid asking the questions, then we will implicitly give our answers by our actions.

To prime our moral intuition, imagine the following scenario: You are a doctor attending to a patient who seems to be in a vegetative state. Their vital signs indicate that they are alive (they are still breathing and have a heart rate), but they show no signs of sentient awareness. They could be brain dead (and not experiencing anything at all), or they could still be sentient, but have no way of communicating it to the world (a real-world condition known as locked-in-syndrome). For the sake of argument, you have no way of distinguishing between these two conditions. You have the opportunity to administer pain medicine. Perhaps the patient was in a terrible car accident, and you know that people in similar accidents experience tremendous pain. Would you administer the pain medicine?

I would give the analgesic: If the patient is brain dead and has no experience, I would be giving pain medicine to someone who was functionally dead, but I would not be making anything worse. However, if the patient is sentient, I could be sparing them excruciating pain. I believe that the possibility of saving a sentient entity from suffering justifies the small risk of wasting the pain medicine. Furthermore, I believe this thought experiment allows us to make a stronger claim than "if an entity is sentient, then it has some moral status": If we believe there is a non-zero probability an entity is sentient, then it has moral status.

This approach can be viewed as an extension of the precautionary principle (articulated by Elder[18]): "Where there is no conclusive consensus, the burden of proof that an action is not harmful falls on the person who is acting...one suspends action that may be potentially harmful until it has been proven harmless". This principle was articulated in the context of fish: If we are uncertain whether fish are sentient, we ought to treat fish as if they were sentient (considering the large amount of suffering that fish might be experiencing).

As long as it is possible that sentience is a substrate independent phenomenon, there is a non-zero probability that computers could be sentient. Consider two scenarios of how we might treat AI systems that may be sentient.

First, suppose that we treat them as if they were sentient when in fact, they are not sentient. This scenario certainly has disadvantages for us: We would be unnecessarily concerned for the experiences of objects which have no capacity for experience. We would put ourselves at a disadvantage by acting over-cautiously, assuming a completely irrelevant moral burden. However, we would not need to be worried that we were committing an ethical transgression.

Now consider the opposite scenario, in which we treat systems as if they were not sentient, and in fact, they are. This could be a catastrophe. If their sentience were analogous to human sentience, how grievously immoral would it be to treat them as if they were objects with no capacity for joy or suffering? We would be neglecting these experiences of billions of entities.

5 Conclusion

With the rapid development of artificial intelligence, we will be in the position of having to make important decisions without a full understanding of relevant facts and issues. Given our lack of understanding combined with the moral catastrophe of potentially causing untold suffering by assuming that a sentient creature is not sentient, I propose the following approach: As long as there is a non-zero probability that a machine is sentient, such a machine should be granted some moral status. Given the magnitude of risk associated with treating sentient computers as if they were merely tools, I see no reasonable justification for assuming and acting as if they will never be sentient.

Acknowledgment

I would like to thank Dr. Ruth Starkman (Stanford University) for her invaluable guidance and expertise.

References

[1] R. Kirk, Sentience and behavior, *Mind*, vol. 83, pp. 43–60, 1974.

[2] T. Nagel, What is it like to be a bat? *Philos. Rev.*, vol. 83, no. 4, pp. 435–450, 1974.

[3] R. Kirk, Zombies, https://plato.stanford.edu/archives/fall2023/entries/zombies/, 2023.

[4] D. Jannai, A. Meron, B. Lenz, Y. Levine, and Y. Shoham, Human or not? A gamified approach to the Turing test, arXiv preprint arXiv: 2305.20010, 2023.

[5] N. Bostrom, Are we living in a computer simulation? *Philos. Q.*, vol. 53, no. 211, pp. 243–255, 2003.

[6] D. J. Chalmers, Absent qualia, fading qualia, dancing qualia, in *Conscious Experience*, T. Metzinger, ed. Exeter, UK: Imprint Academic, 1995, pp. 309–328.

[7] G. Tononi, An information integration theory of consciousness, *BMC Neurosci.*, vol. 5, p. 42, 2004.

[8] R. Penrose, *Shadows of the Mind*. Oxford, UK: Oxford University Press, 1994.

[9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, Language models are unsupervised multitask learners, *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[11] OpenAI, GPT-4 technical report, Technical report, OpenAI, San Francisco, CA, USA, 2023.

[12] C. Biever, ChatGPT broke the Turing test—The race is on for new ways to assess AI, *Nature*, vol. 619, no. 7971, pp. 686–689, 2023.

[13] S. D. Baum, B. Goertzel, and T. G. Goertzel, How long until human-level AI? Results from an expert assessment, *Technol. Forecast. Soc. Change*, vol. 78, no. 1, pp. 185–195, 2011.

[14] D. J. Chalmers, Facing up to the problem of consciousness, *Journal of Consciousness Studies*, vol. 2, no. 3, pp. 200–219, 1995.

[15] C. McGinn, Can we solve the mind-body problem? *Mind*, vol. 99, no. 391, pp. 349–366, 1989.

[16] D. Bourget and D. J. Chalmers, Philosophers on philosophy: The 2020 PhilPapers survey, *Philos. Impr.*, vol. 23, no. 1, p. 11, 2023.

[17] R. D. Ryder, Souls and sentientism, *Between the Species*, vol. 7, no. 1, p. 3, 1991.

[18] M. P. Elder, The fish pain debate: Broadening humanity's moral horizon, *J. Anim. Ethics*, vol. 4, no. 2, pp. 16–29, 2014.

Luke R. Hansen received the bachelor degree in symbolic systems from Stanford University, USA in 2023. He is currently pursuing the master degree in computer science (artificial intelligence specialization) at Stanford University, USA. He is interested in trying to understand the mind in computational terms and leveraging insights from cognitive neuroscience to improve artificial intelligence algorithms