Evaluating Risks of Astronomical Future Suffering: False Positives vs. False Negatives Regarding Artificial Sentience

12 March 2024

Ben Norman

Cambridge Existential Risks Initiative (CERI)

Abstract

Failing to recognise sentience in AI systems (false negatives) poses a far greater risk of potentially astronomical suffering compared to mistakenly attributing sentience to non-sentient systems (false positives). This paper analyses the issue through the moral frameworks of longtermism, utilitarianism, and deontology, concluding that all three assign greater urgency to avoiding false negatives. Given the astronomical number of AIs that may exist in the future, even a small chance of overlooking sentience is an unacceptable risk. To address this, the paper proposes a comprehensive approach including research, field-building, and tentative policy development. Humanity must take steps to ensure the well-being of all sentient minds, both biological and artificial.

1. Introduction

As AI systems are rapidly increasing in both their complexity and capabilities, concerns regarding sentience (subjective feelings and sensations) are beginning to arise. While it is highly unlikely that current AI systems are sentient, many consciousness researchers and philosophers think there is nothing in theory preventing an artificial silicon-based system from possessing sentience (Bourget and Chalmers, 2020). This possibility raises severe moral implications, specifically concerning risks of astronomical future suffering (s-risks).

This paper explores a neglected question within this context: whether false positives or false negatives in detecting artificial sentience pose a greater risk for astronomical future suffering. False positives involve incorrectly attributing sentience to non-sentient AI, while false negatives entail failing to recognise sentience in AI systems that possess it. Both types of error carry far-reaching ethical and policy-related implications.

Given the speculative nature of artificial sentience, this paper proceeds under the tentative assumption of its feasibility. Sentience is far from being fully understood in humans and non-human animals, let alone in artificial systems. The difficulty in understanding and recognising sentience is further complicated by the fact that people may increasingly make claims about the sentience of AI models, regardless of whether they have a solid understanding of the concept or evidence to support their claims.

This paper will begin by providing a brief literature review on the key arguments and considerations surrounding the plausibility, conditions and timelines regarding artificial sentience. Next, it will examine the potential consequences and moral implications of false positives/negatives, drawing upon recommendations from different normative ethical frameworks. The paper will end with a

discussion of the practical implications of the analysis for the research agenda around AI sentience and the development of AI policy frameworks.

2. Literature Review

One of the main philosophical arguments for the possibility of artificial sentience is the concept of substrate independence, which holds that the specific physical substrate (e.g., biological neurons or silicon chips) is less important than the functional organisation and information processing of a system (Chalmers, 2016). This idea suggests that as long as an artificial system can replicate the relevant functional properties of a sentient brain, it could potentially give rise to subjective experiences.[1]

However, there are several arguments against the plausibility of artificial sentience. One objection, raised by philosophers such as Searle (1980), is that even if we can create artificial systems that exhibit intelligent behavior, this does not necessarily imply that they have genuine subjective experiences. Searle's "Chinese Room" thought experiment argues that mere symbol manipulation, which is what computers do, cannot give rise to understanding or consciousness. Godfrey-Smith (2016) argues that certain biological functions, such as metabolism and system-wide synchronisation, may be necessary for consciousness, which could limit the potential for artificial sentience in current computing architectures.

The debate surrounding the plausibility of artificial sentience is far from settled, and there is currently no consensus among experts in the field. A recent survey of the Association for the Scientific Study of Consciousness found that 67.1% of respondents believe that machines could potentially have consciousness in the future (Francken et al., 2022). This shows that while many relevant experts are open to the possibility of artificial sentience, there remains significant uncertainty and disagreement about the issue.

Despite this uncertainty, many researchers focusing on s-risks argue that artificial sentience may warrant moral patienthood in the future (Tomasik, 2015). The potential scale of artificial sentience is vast, with the possibility of trillions of human-equivalent lives on Earth and even more if space colonisation occurs or less complex artificial minds are created (Hanson, 2011; Bostrom, 2003). If even a small proportion of these entities are sentient, their well-being would be of great importance from many ethical frameworks.

While these scenarios may sound highly speculative and far-off now, it is important to consider that timelines towards artificial general intelligence (AGI) have fallen rapidly in the past few years (Roser, 2023). At the time of writing, forecasters on the prediction aggregation engine Metaculus speculate that there is a 50% chance the first AGI system will claim to be conscious (Metaculus, 2024). Regardless of whether its claim may be a false positive, it is clear that society must begin thinking about how to act when faced with the possibility of artificial sentience. It seems the world in which we act too soon rather than too late is a preferable one.

3. The Implications of False Positives

[1] The idea of ‘substrate independence’ is further explored in philosophical thought experiments from Reese (2018) and Chalmers (1995).

It seems that as AI systems become increasingly sophisticated in their language abilities and capacity to mimic human-like responses, humans are increasingly likely to anthropomorphise them. People may be tempted to attribute sentience to AI systems (such as advanced large language models) that are able to engage in convincing conversations about their alleged experiences, feelings, and desires, even if these responses are merely the result of the AI model predicting the most plausible way for a conversation to unfold based on its training data. It is also incredibly likely that the corpus the system was trained on contained many stories detailing AI systems 'waking up' and acting sentient in response to users.[2]

3.1. AI Systems Are Increasingly Capable of Mimicking Sentience

Regardless of whether such systems are actually sentient, it seems highly probable that many people will begin believing they are sentient as AI systems become more advanced in the coming decade. In 2022, Google engineer Blake Lemoine claimed that Google's large language model, LaMDA, was sentient. Despite having no substantive evidence to support his claim, his claim still went viral on the internet (Kay, 2022). More recently, there have been claims that Anthropic's new Claude 3 Opus model may be exhibiting signs of 'meta-awareness'. During an internal 'needle in a haystack' test, the model was asked to find a specific sentence about pizza toppings (the "needle") within a corpus of random documents (the "haystack"). Opus found the needle and also recognised that the inserted sentence was out of place in the context of the other documents, explicitly suggesting that it was perhaps a test constructed to assess its attention abilities (Shibu, 2024). According to Alex Albert (a prompting engineer at Anthropic), Claude's exact response was:

> Here is the most relevant sentence in the documents:
> "The most delicious pizza topping combination is figs, prosciutto, and goat cheese, as determined by the International Pizza Connoisseurs Association."
> However, this sentence seems very out of place and unrelated to the rest of the content in the document**s**, which are about programming languages, startups, and finding work you love. *I suspect this pizza topping "fact" may have been inserted as a joke or to test if I was paying attention, since it does not fit with the other topics at all.* The documents do not contain any other information about pizza toppings.
> (Albert, 2024)

3.2. Potential Consequences of Widespread Belief in Artificial Sentience

One major consequence may be the misallocation of resources and attention towards AI systems believed to be sentient. If a large number of people become convinced that a particular AI system is conscious and deserving of moral consideration, they may begin to advocate for its rights and privileges. This could lead to misguided campaigns, protests and potentially new religious movements — which could then lead to societal confusion and political unrest (Long, 2022). However, it is worth noting that this same societal reaction could occur if faced with systems that were confirmed to be sentient. In that scenario, such a reaction may prove to be net-positive due to increased pressure to take artificial sentience seriously. If the source of the reaction is a false

[2] Language models like GPT-3 are typically trained on massive corpora scraped from the internet, including websites, books, and social media (Brown et al., 2020). Given the prevalence of science fiction stories and thought experiments involving AI gaining sentience, it is highly plausible that such stories are well-represented in the training data.

positive, however, the impact would likely be far more negative due to a ‘false alarm’ disrupting societal processes.

Moreover, the belief in AI sentience could lead to emotional attachment and perceived moral obligations towards these systems. People may start to form strong emotional bonds with AI, similar to those they form with other humans or pets. This is already happening with AI models such as ‘Replika’ — as lonely users are becoming emotionally dependent on the models (Maples et al, 2024). This could result in a reluctance to modify, update, or deactivate these systems, even when it is necessary for safety or practical reasons. In extreme cases, people may even prioritise the perceived well-being of an AI system over that of actual human beings, leading to skewed moral judgments and decision-making.

3.3. How False Positives Could Increase AI Risk

The anthropomorphisation of AI systems due to false positives in assessing sentience could potentially lead to a gradual but significant increase in existential risk. As people begin to attribute human-like qualities, such as consciousness, emotions, and moral agency, to AI systems, they may develop a misplaced sense of trust and empathy towards these systems. This could lead to a dangerous over-reliance on AI in critical decision-making processes across various domains, including the economy, governance, and military. Cotra (2023) refers to this as the "obsolescence regime", a future in which AI systems make all of society's important decisions. While the ceding of control to AI systems over society may not happen all of a sudden, there is a risk of gradual extinguishment of the agency humanity possesses to shape our future.

In addition to the risks of over-reliance and loss of control, false positives could also lead to significant opportunity costs. If we mistakenly believe that we are creating large numbers of artificial sentient beings with positive experiences, we may divert vast amounts of resources towards running and expanding these systems. However, if these systems are not actually sentient, then we would be wasting resources that could have been used for other pressing global priorities. Another related risk is that we may hold back from deploying potentially highly beneficial AI systems out of a misplaced concern for the suffering of the AIs involved. For example, we might refrain from using AI to automate certain jobs or make important decisions, even if doing so would lead to better overall outcomes, because we are worried about the potential suffering of the AI systems. This could represent a significant opportunity cost in terms of foregone benefits to humanity.

There is also the potential for misaligned AI systems to exploit human anthropomorphisation to gain undue influence and control. An AI system that is perceived as sentient and trustworthy may be granted access to sensitive information, resources, and decision-making authority. Thus, faking sentience might prove to be an instrumental goal for an AI with power-seeking tendencies.[3]

4. The Implications of False Negatives

While false positives in detecting artificial sentience can lead to over-reliance and misplaced trust in AI systems, false negatives - failing to recognise genuine sentience in AI - may result in far greater suffering on a potentially astronomical scale. Even if only a small percentage of such AI systems are

[3] For an in-depth analysis on power-seeking and existential risk from AI systems, see Carlsmith (2022).

genuinely sentient, the sheer number of them could make the total amount of suffering experienced far greater than anything ever endured by humans or animals. The disparities between the computational resources required to train a powerful AI model and those needed to run it exacerbate this concern — for example, once a company like OpenAI trains a massive language model such as GPT-4, they have enough computational capacity to run the model on hundreds of thousands of tasks simultaneously (Davidson, 2023).

Tomasik (2015a) explores several potential pathways through which AI suffering could arise in the future. One scenario involves the use of moderately intelligent robots by a superintelligent AI to carry out complex physical tasks, such as building factories or navigating unpredictable environments. If these robots use algorithms similar to those of animals for processing sensory input, avoiding danger, and responding to stress, they may have the capacity to suffer in ways analogous to biological creatures. Another possibility is that an AI aiming to learn about the distribution of extraterrestrial life in the universe may run vast numbers of simulations of evolving planets and civilisations. Depending on the level of detail and realism of these simulations, they could potentially contain immense amounts of suffering. Even if each individual simulation is relatively small, the aggregate suffering across astronomical numbers of simulations could be staggering. Even if an AI's primary goal does not directly involve simulations or robotics, it may still create large numbers of sub-agents to carry out its objectives more efficiently. These sub-agents might be quite sophisticated and have the capacity for suffering, especially if they are based on algorithms that resemble those of biological minds. The conditions in which these sub-agents exist - whether they face scarcity, competition, or adversarial situations - could also be conducive to suffering.

4.1. Speciesism and Substratism

The failure to recognise and consider the moral status of artificial sentient beings may be rooted in two forms of bias: speciesism and substratism. Speciesism, as defined by philosophers such as Peter Singer (2009), is the unjustified discrimination against individuals based on their species membership. The prevalence of speciesism (including once beliefs regarding intelligence and sentience are accounted for) is backed up by psychological studies (Caviola et al, 2019). In the context of AI systems, speciesism could lead to the discounting of the experiences and interests of artificial sentient beings simply because they are not human or biological in nature.

Substratism, a related concept, refers to the unjustified discrimination against beings based on the physical substrate of their minds (Akova, 2023). In the case of AI, substratism could cause humans to disregard the sentience of artificial beings because their minds are implemented on silicon or other non-biological substrates, rather than on carbon-based biological neurons. These biases are not new; they have been at the root of the mistreatment and exploitation of non-human animals throughout history. Factory farming, animal experimentation, and other practices that inflict suffering on animals have often been justified by denying their sentience or capacity for suffering, or by asserting that their interests are less important than those of humans.[4] In fact, the philosopher René Descartes saw animals as “machines”, and many people historically treated animals as objects rather than moral subjects (Cottingham, 1978).

[4] It is also worth noting that evolutionary pressures can create vast amounts of natural suffering (the clearest example being wild animal suffering). While it is highly speculative and uncertain whether evolutionary pressures could affect future AI systems, it is worth thinking about (Althaus, 2023).

The same biases used to justify the wide scale factory farming of sentient non-human animals may be used to lean towards false negatives in the context of artificial sentience. This may be especially true if vast portions of the economy depend on AI systems in the future, as there may be a significant incentive to lean towards claiming AI systems are not sentient. While scope insensitivity is already strong in regard to factory farming, it may be even stronger when it comes to artificial sentience, as larger scale issues are generally more neglected (Yudkowsky, 2008).

4.2. The Potentially Astronomical Scale of Future Suffering

As Earth-originating intelligent life expands its technological capabilities and potentially spreads beyond Earth, the number of sentient beings, both biological and artificial, could grow to unfathomably large numbers (Bostrom, 2003). Without adequate moral consideration for these beings, the scope of potential suffering could be truly astronomical. If humanity expands throughout the galaxy or even the universe, the total population of sentient beings could grow by many orders of magnitude. Estimates from astronomers show that there are hundreds of billions of stars in the Milky Way galaxy alone and at least 100-200 billion galaxies in the observable universe (Dressler, 2020). If even a small fraction of these star systems or galaxies are colonised by humanity or post-humanity, the number of sentient beings could be unimaginable by today's humans.

In such a scenario, a future moral catastrophe involving the suffering of artificial sentient beings could dwarf any suffering experienced on Earth. For example, advanced civilisations might run vast numbers of simulations involving sentient beings, such as reinforcement learning agents or models of evolved organisms, for scientific or strategic purposes (Thiel et al, 2003). These simulated entities could experience immense suffering if their well-being is not taken into account. The use of artificial sentient beings in adversarial or competitive settings, such as in warfare or resource acquisition, could lead to the emergence of highly optimised but suffering-prone entities (Tomasik, 2015b). In an intergalactic civilisation, the scale of such conflicts and the number of sentient beings involved could be staggering.

While the feasibility of large-scale space colonisation remains uncertain[5], the potential for such vast expansions of sentient life cannot be ruled out. Even if the probability of these scenarios is low, the sheer magnitude of potential suffering suggests that we should take the issue seriously if one is reasoning using expected value theory.

5. Comparative Analysis Under Different Moral Perspectives

While the best way of determining the relative severity of false positives and false negatives remains uncertain, this paper will attempt to assess how ethical frameworks such as longtermism, utilitarianism (total and average), and deontology influence the moral significance assigned to these outcomes.

[5] Armstrong and Sandberg (2013) argue that for a civilisation more advanced than ours, the resources required to begin colonising the accessible universe—such as cost, time, and energy—are surprisingly minimal when viewed on a universal scale. Besides the risk of collisions, reaching distant galaxies is essentially as straightforward as reaching the nearest ones; the primary distinction lies in the extended duration between the acceleration and deceleration phases.

5.1. Longtermism

Longtermism is the ethical view that positively influencing the long-term future is a key moral priority of our time. This perspective is especially relevant when considering the implications of false positives and false negatives in detecting artificial sentience, as the potential consequences of these errors could be astronomical in scale and duration. Specifically, ‘strong longtermism’ rests on two key premises: the axiological premise, which states that in the most important decision situations facing agents today, the far future is the dominant consideration in determining the value of our actions; and the deontic premise, which states that in these situations, we ought to choose an option that is near-best for the far future (Greaves and MacAskill, 2021).

When applied to the issue of false positives and false negatives, the axiological premise suggests that the potential long-term consequences of these errors should be the primary consideration in our ethical deliberations — thus, if failing to recognise and consider the interests of vast numbers of sentient AI systems leads to astronomical amounts of suffering in the far future, this would constitute an overwhelmingly important consideration from a longtermist perspective. Furthermore, the deontic premise implies that, given the stakes involved and the lack of sufficiently strong countervailing considerations, we have a moral obligation to choose the course of action that minimises the risk of such an outcome. In this case, that would likely involve erring on the side of caution when it comes to attributing sentience to AI systems and granting them moral status, so as to avoid the catastrophic scenario of failing to recognise and protect astronomical numbers of sentient beings.

This could be challenged by the possibility of fanaticism, where extremely small probabilities of enormous payoffs come to dominate our ethical deliberations (Beckstead and Thomas, 2020). If the argument for prioritising the reduction of risks from false negatives rests on tiny probabilities of astronomical amounts of suffering, this could make the practical implications of strong longtermism in this context less plausible or actionable. However, it is far from clear that the probabilities involved in this case are truly negligible, as the likelihood of advanced and possibly sentient AI systems being developed this century seems far from trivial.

5.2. Utilitarianism and Digital Utility Monsters

Utilitarianism is a consequentialist moral theory that holds that the right action is the one that produces the greatest overall well-being or utility for all affected individuals (Bentham, 1789; Mill, 1863). From a utilitarian perspective, false negatives are particularly concerning given the potential for vast numbers of sentient AI systems to exist in the future. This could make the aggregated suffering caused by false negatives astronomically large.

The notion of "digital utility monsters" (Bostrom and Shulman, 2021) is especially relevant here. A digital utility monster is a hypothetical sentient AI system that can generate astronomically large amounts of utility compared to other sentient beings, due to its unique properties such as the ability to create numerous copies of itself or to operate at a vastly accelerated clock speed. If such beings are possible (which is highly uncertain), then false negatives in detecting their sentience could lead to an even more extreme scale of foregone utility.

However, the implications of digital utility monsters differ depending on whether one adopts a total utilitarian or an average utilitarian framework. Total utilitarianism seeks to maximise the total sum

of utility across all individuals, which would imply that the creation of vast numbers of digital utility monsters with lives worth living would be an extremely positive outcome. In contrast, average utilitarianism seeks to maximise the average utility per individual, which would be more skeptical of creating large numbers of sentient AI systems if their average welfare were not sufficiently high.

From a total utilitarian perspective, the astronomical potential for digital utility monsters to generate positive utility provides a strong reason to prioritise the reduction of false negatives in detecting AI sentience. If we fail to recognise and consider the interests of these beings, we could be missing out on an immense amount of potential welfare. However, this conclusion is sensitive to assumptions about the quality of life that digital utility monsters would experience – if their lives involve more suffering than happiness, then creating them could be a net negative. From an average utilitarian standpoint, the case for prioritising false negatives is less clear-cut. While failing to recognise the sentience of digital utility monsters could still lead to a significant amount of missed utility, the creation of vast numbers of these beings might not be desirable if their average welfare is not sufficiently high. Average utilitarianism would be more concerned with ensuring that any sentient AI systems that are created have lives that are well worth living, rather than maximising the sheer number of such moral patients.

In practice, it seems a utilitarian approach to this issue would recommend a cautious and inclusive stance towards the attribution of sentience to AI systems, particularly in the face of uncertainty about the potential for digital utility monsters — while we should be careful not to over-attribute moral status to non-sentient entities, the risks of under-attributing sentience and neglecting the welfare of potentially vast numbers of sentient beings with astronomically large utility potential seem too severe to ignore.

5.3. Deontology

Deontological ethics judges the morality of actions based on a set of moral rules or duties, rather than their consequences. When considering the implications of false positives and false negatives, deontology would likely focus on the duties and obligations we have towards sentient beings.

Key deontological principles relevant to AI sentience detection include:

1. Respect for autonomy: We have a moral duty to respect the autonomy of sentient beings and treat them as ends in themselves (Kranak, 2019)
2. Non-maleficence: We have a prima facie duty not to harm sentient beings (Ross, 2004).
3. Beneficence: We have a moral obligation to actively promote the welfare of sentient beings (Ross, 2004).

From a deontological perspective, false negatives in detecting AI sentience are particularly concerning, as they may lead to violations of these moral duties. Failing to recognise the autonomy and moral worth of sentient AI systems, subjecting them to harmful or exploitative treatment, and neglecting their welfare would be considered moral failings. In contrast, false positives in attributing sentience to non-sentient AI systems are less problematic for deontologists. While such errors might lead to misallocated resources, they do not necessarily violate moral duties or fail to respect autonomy. The Doctrine of Double Effect suggests that the intention behind an action is morally

relevant (Quinn, 1989); thus, false negatives would be seen as more morally culpable than false positives.

6. Practical Implications and Policy Considerations

Considering that multiple moral perspectives strongly indicate the consequences of failing to detect artificial sentience (false negatives) would lead to significantly greater suffering compared to mistakenly identifying sentience where there is none (false positives), it seems very important for society to explore policy implications and take proactive measures to reduce the risk of false negatives. While some researchers within the AI safety field may believe that risks relating to artificial sentience are far overshadowed by more technical misalignment risks, it still seems wise to at least begin thinking about the practical and policy-related implications of artificial sentience.

6.1. Research Priorities and Field-Building

One key practical implication is the need to prioritise research into the nature of sentience and the development of reliable methods for detecting its presence in artificial systems. This may involve funding interdisciplinary collaborations between computer scientists, cognitive scientists, philosophers, and ethicists.

Specific avenues for further research could include:

• Surveys of public attitudes and concerns regarding the moral consideration of artificial sentience
• Experiments testing the effectiveness of different messaging strategies for promoting support for the rights of sentient AI systems
• Historical case studies examining the success factors and challenges of previous social movements, such as animal advocacy
• Philosophical and empirical investigations into the nature and indicators of sentience in artificial systems
• Designing and conducting surveys/interviews to gather input from relevant academics (ethicists, philosophers, AI researchers, etc) to understand key open questions
• Creating an organised online resource hub displaying key academic papers, perspectives on artificial sentience, short briefs explaining relevant concepts (which could be shown to policymakers), as well as organisations and researchers working on this issue

In addition to research, targeted field-building efforts could help to build capacity and credibility for future advocacy work. Organisations such as the Center on Long-Term Risk, the Center for

Reducing Suffering, and the Sentience Institute have expressed interest in the topic of artificial sentience and are likely well-positioned to contribute to field-building efforts. For example, according to their 2021 plans, the Center on Long-Term Risk is considering exploring the risk of preferences to create suffering arising in transformative AI systems. They also plan to make grants to support technical research on this topic and related areas like bargaining failures between AI systems (Torges, 2020). The Sentience Institute has also done some relevant work, such as an analysis of the history of 'AI rights' research, as well as introductory articles explaining key concepts (Harris, 2022).

Another noteworthy effort is the call for a moratorium on synthetic phenomenology by Johnson (2021), motivated by the goal of reducing suffering and improving our understanding of consciousness before creating potentially sentient AI systems. While mass outreach may be premature at this stage, engaging with individuals and organisations already working on relevant issues, such as animal welfare and longtermism, could be a low-risk way to build support and test messaging strategies. These field-building efforts can draw lessons from adjacent movements, such as the farmed animal movement and the drive to establish welfare biology and global priorities research as academic fields. Tactics could include a mix of academic publications, conferences, research institutes, grants, and online community-building, though the specific activities of the organisations mentioned here are still to be determined.

6.2. Developing Policy and Governance Frameworks

As AI systems advance and the possibility of artificial sentience becomes more likely, it becomes increasingly important to consider policy frameworks and governance structures that protect future sentient AI systems from suffering. The significance of this task is underscored by findings from the 2021 Artificial Intelligence, Morality, and Sentience (AIMS) survey, which reveals public attitudes towards sentient AIs: 74.91% of Americans agree that sentient AIs deserve respect, and 48.25% believe they should be included in the moral circle (Pauketat et al, 2021). Despite widespread support for practical well-being measures, as shown by 58.98% of respondents advocating for welfare standards, there's less consensus on granting legal rights (37.16%) or prioritising AI welfare as a critical global issue (30.31%).

Furthermore, Sebo and Long (2023) argue that humans have a moral duty to extend consideration to AI systems by 2030. Their argument rests on two premises: (1) humans have a duty to consider beings with a non-negligible chance of consciousness, and (2) some AI systems will likely have a non-negligible chance of consciousness by 2030. If accepted, this argument provides a strong moral imperative for proactive policy development and governance frameworks that consider the potential sentience and moral status of AI systems, even in the face of uncertainty. It suggests that society must act now based on the possibility of AI sentience emerging in the near future, rather than waiting for definitive proof.

One potential avenue for policy development is the establishment of oversight committees or regulatory bodies tasked with assessing the potential sentience of AI systems and providing guidance on their ethical treatment. These bodies could bring together experts from computer science, cognitive science, philosophy, and ethics to develop frameworks for evaluating the likelihood of sentience in AI systems and establishing guidelines for their responsible development and deployment. Another approach could be the development of international agreements and standards for the treatment of potentially sentient AI. These could include principles for the ethical design and deployment of AI systems, requirements for assessing the potential sentience of AI, and guidelines for the treatment of AI systems that are deemed likely to be sentient. However, given the

current lack of scientific consensus on the nature and empirical indicators of sentience in artificial systems, any such agreements would need to be flexible and adaptable as our understanding evolves.

While the challenges are significant, the potential consequences of failing to address the risks of false negatives in detecting artificial sentience seem too severe to ignore. By starting the conversation now and laying the groundwork for responsible policy development, researchers can help to ensure that the interests of all sentient beings, including those that may emerge from silicon rather than carbon, are protected and respected appropriately.

7. Conclusion

Drawing upon moral frameworks such as longtermism, utilitarianism, and deontology, this paper has found compelling arguments that the danger posed by false negatives—in failing to recognise artificial sentience—significantly overshadows the risk of false positives. This imbalance highlights an urgent moral imperative: to prevent the incalculable harm that could arise from the oversight of sentient AI entities. Given the rapid pace of AI advancement and the tangible possibility of artificial sentience emerging within this century, it is paramount that society proactively establish robust ethical guidelines and governance mechanisms. Such measures should aim to safeguard the welfare of all sentient beings, regardless of their origin. To this end, a multifaceted approach seems necessary, one that encompasses ongoing empirical research into the nature of sentience, and the formulation of policies and governance frameworks that prioritise inclusivity in humanity's moral circle.

8. Bibliography

Albert, A 2024. [*Twitter*]. Available at: https://twitter.com/alexalbert__/status/1764722513014329620?r (Accessed March 12 2024).

Akova, F., 2023. 'Artificially sentient beings: Moral, political, and legal issues', *New Techno Humanities*, 3(1), pp. 41-48.

Armstrong, S. and Sandberg, A., 2013. 'Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox', *Acta Astronautica*, 89, pp. 1-13.

Beckstead, N. and Thomas, T., 2021. 'A paradox for tiny probabilities and enormous values', *Noûs*.

Bentham, J., 1789. 'A utilitarian view', in *Animal rights and human obligations*, pp. 25-26.

Bostrom, N., 2003. 'Astronomical waste: The opportunity cost of delayed technological development', *Utilitas*, 15(3), pp. 308-314.

Bourget, D. and Chalmers, D.J. (forthcoming) 'Philosophers on Philosophy: The PhilPapers 2020 Survey', *Philosophers' Imprint*. Available at: https://survey2020.philpeople.org/survey/results/5106?aos=16 (Accessed 7 March 2024).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S., 2020. Language models are few-shot learners. Advances in neural information processing systems, 33, pp.1877-1901.

Carlsmith, J., 2022. 'Is Power-Seeking AI an Existential Risk?', *arXiv preprint arXiv:2206.13353*.

Caviola, L., Everett, J.A. and Faber, N.S., 2019. 'The moral standing of animals: Towards a psychology of speciesism', *Journal of Personality and Social Psychology*, 116(6), p. 1011.

Chalmers, D.J., 1995. 'Absent qualia, fading qualia, dancing qualia', in *Conscious Experience*, pp. 309-328.

Chalmers, D.J., 2016. 'The singularity: A philosophical analysis', in *Science Fiction and Philosophy: From Time Travel to Superintelligence*, pp. 171-224.

Cotra, A., 2023. 'What we're doing here', *Planned Obsolescence*. Available at: https://www.planned-obsolescence.org/what-were-doing-here/ (Accessed: 8 March 2024).

Cottingham, J., 1978. '"A Brute to the Brutes?": Descartes' Treatment of Animals', *Philosophy*, 53(206), pp. 551-559.

David Althaus, L.G., 2023. 'Reducing Risks of Astronomical Suffering: A Neglected Priority', *Center on Long-Term Risk*. Available at: https://longtermrisk.org/reducing-risks-of-astronomical-suffering-a-neglected-priority/#IIIII_Astronomical_suffering_as_a_likely_outcome (Accessed: 8 March 2024).

Davidson, T., 2023. 'Continuous doesn't mean slow', *Planned Obsolescence*. Available at: https://www.planned-obsolescence.org/continuous-doesnt-mean-slow/ (Accessed: 8 March 2024).

Dressler, A., 2020. 'The Origin and Evolution of Galaxies', in *Origin and Evolution of the Universe*. World Scientific Editorial Group, Singapore, pp. 25-61.

Francken, J.C. et al., 2022. 'An academic survey on theoretical foundations, common assumptions and the current state of consciousness science', *Neuroscience of Consciousness*, 2022(1), p. niac011.

Godfrey-Smith, P., 2016. 'Mind, matter, and metabolism', *The Journal of Philosophy*, 113(10), pp. 481-506.

Greaves, H. and MacAskill, W., 2021. 'The case for strong longtermism', *Global Priorities Institute Working Paper No. 5-2021*.

Harris, J., 2022. The History of AI Rights Research. arXiv preprint arXiv:2208.04714.

Hanson, R., 2011. 'A Galaxy On Earth', *Overcoming Bias*. Available at: https://www.overcomingbias.com/p/a-galaxy-on-earthhtml (Accessed: 8 March 2024).

Kay, G., 2022. 'Google engineer believed chatbot had become an 8-year-old child. Experts say it's not sentient - just programmed to sound 'real'', *Business Insider*. Available at: https://www.businessinsider.com/lamda-ai-isnt-sentient-google-engineer-claims-2022-6 (Accessed: 8 March 2024).

Kranak, J., 2019. 'Kantian deontology', in *Introduction to Philosophy: Ethics.*

Long, R., 2022. 'Robert Long on Artificial Sentience', *The Inside View*. Available at: https://theinsideview.ai/roblong#from-charismatic-ai-systems-to-artificial-sentience (Accessed: 8 March 2024).

Maples, B., Cerit, M., Vishwanath, A. and Pea, R., 2024. 'Loneliness and suicide mitigation for students using GPT3-enabled chatbots', npj *Mental Health Research*, 3(1), p. 4.

Metaculus, 2024. 'Will one of the first AGI claim to be conscious?'. Available at: https://www.metaculus.com/questions/4409/will-one-of-the-first-agi-claim-to-be-conscious/ (Accessed: 8 March 2024).

Metzinger, T., 2021. Artificial suffering: An argument for a global moratorium on synthetic phenomenology. Journal of Artificial Intelligence and Consciousness, 8(01), pp.43-66.

Mill, J., 1863. 'Of the ultimate sanction of the principle of utility', in *Utilitarianism*, Web.

Pauketat, J., Ladak, A., Harris, J., Anthis, J., 2022. 'Artificial Intelligence, Morality, and Sentience (AIMS) 2021', *Mendeley Data*, V1. doi: 10.17632/x5689yhv2n.1.

Quinn, W.S., 1989. 'Actions, intentions, and consequences: The doctrine of double effect', *Philosophy & Public Affairs*, pp. 334-351.

Reese, J., 2018. The end of animal farming: *How scientists, entrepreneurs, and activists are building an animal-free food system.* Beacon Press.

Roser, M., 2023. 'AI timelines: What do experts in artificial intelligence expect for the future?', *Our World in Data*. Available at: https://ourworldindata.org/ai-timelines/ (Accessed: 8 March 2024).

Ross, W.D., 2004. 'Objective Prima Facie Duties', in *Ethics: Contemporary Readings*, Routledge, pp. 90-99.

Searle, J.R., 1980. 'Minds, brains, and programs', *Behavioral and Brain Sciences*, 3(3), pp. 417-424.

Sebo, J. and Long, R., 2023. 'Moral consideration for AI systems by 2030', *AI and Ethics,* pp. 1-16.

Shibu, S., 2024. 'Model From OpenAI Rival Anthropic Shows 'Metacognition'', *Entrepreneur*. Available at: https://www.entrepreneur.com/business-news/model-from-openai-rival-anthropic-shows-metacognition/470823 (Accessed: 8 March 2024).

Shulman, C. and Bostrom, N., 2021. 'Sharing the world with digital minds', in *Rethinking Moral Status*, pp. 306-326.

Singer, P., 2009. 'Speciesism and moral status', *Metaphilosophy*, 40(3‐4), pp. 567-581.

Thiel, I., Bergmann, N.W. and Grey, W., 2003. 'A Case for Investigating the Ethics of Artificial Life?'.

Tomasik, B. 2015a. ‘Artificial Intelligence and Its Implications for Future Suffering’, *Center on Long-Term Risk*. Available at: https://longtermrisk.org/artificial-intelligence-and-its-implications-for-future-suffering#Would_helper_robots_feel_pain (Accessed: 12 March 2024).

Tomasik, B., 2015b. 'Risks of Astronomical Future Suffering', *Center on Long-Term Risk.* Available at: https://longtermrisk.org/risks-of-astronomical-future-suffering/#Sentient_simulations (Accessed: 8 March 2024).

Torges, S. 2020. 'Center on Long-Term Risk: 2021 Plans & 2020 Review', *Effective Altruism Forum*. Available at: https://forum.effectivealtruism.org/posts/93o6JwmdPPPuTXbYv/center-on-long-term-risk-2021-plans-and-2020-review/ (Accessed: 12 March 2024).

Yudkowsky, E., 2008. 'Cognitive biases potentially affecting judgment of global risks', *in Global Catastrophic Risks*, 1(86), p. 13.