The ML Community Must Prepare for AI Sentience

Lucius Caviola
University of Oxford
Oxford, UK
lucius.caviola@philosophy.ox.ac.uk

Jeff Sebo
New York University
New York, NY, USA
jeff.sebo@nyu.edu

Sören Mindermann
Mila, Quebec AI Institute
Montreal, Canada
soeren.mindermann@mila.quebec

Abstract

We argue that the machine learning (ML) community must begin preparing for the possibility of AI sentience for two reasons. First, leading scientific theories coupled with emerging architectural evidence support a realistic possibility that future AI systems may develop sentience—the capacity for positive and negative subjective experiences. Even if uncertainty is high and the chances may be low, the moral stakes are high, and precautionary attention is warranted. Second, regardless of whether AI sentience is ever achieved, public belief in AI sentience is likely to grow as systems become more advanced, human-like, and engaging. Such beliefs will shape consumer behavior, policy, and the trajectory of AI development. The ML community must therefore prepare both for the technical possibility of AI sentience and for the societal consequences of widespread belief in AI sentience. We outline reasons for taking this issue seriously, explore drivers of public perception, and recommend concrete steps that ML researchers and institutions can take to address the ethical and societal implications of AI sentience.

1 Introduction

AI systems are advancing at an extraordinary pace. As they become more capable, expressive, and interactive, they increasingly resemble human-like agents—beings with goals, emotions, and personalities. This raises profound questions about their sentience and moral status. Could some AI systems eventually become sentient, capable of positive or negative subjective experience, including happiness and suffering? And if so, what responsibilities would that entail? **This paper argues that the ML community must begin preparing for the possibility that AI sentience—whether real or perceived—will become a significant societal issue.**

First, ML researchers should prepare for AI sentience because near-future AI sentience is a realistic possibility. Scientists and philosophers have started to examine AI systems for evidence of sentience, looking past potentially misleading behavioral evidence for underlying architectural or computational evidence of features associated with subjective experience. This research indicates that while current AI systems possess relatively few features associated with subjective experience, there are no clear technical barriers towards the creation of AI systems with many such features in the near future [8, 18].

Second, ML researchers should prepare for AI sentience because even if AI systems do not in fact become sentient, there may still be significant disagreement and uncertainty about this issue, both among experts and among the general public. People already vary in their judgments about AI

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025).

consciousness and moral status, and as AI systems become more human-like in their linguistic behaviors, physical behaviors, and emotional expressiveness, the belief that these systems have their own thoughts and feelings is likely to grow [17, 14]. These beliefs will influence consumer behavior, public norms, political demands, and regulatory frameworks, regardless of accuracy [16, 9].

This paper argues that ML researchers are not mere passive observers of these technical and societal issues—they play an active role in shaping how AI systems are built and perceived, and they have a responsibility to prepare to play this role well. Sections 2 and 3 examine the technical and societal cases for taking AI sentience seriously within ML. Section 4 then provides concrete recommendations for ML researchers and institutions. If the ML community fails to prepare, it may be caught flat-footed—ethically, scientifically, and politically—when the question of AI sentience moves from speculative debate to societal crisis.

2 AI Sentience Is a Realistic Possibility

Part of why the ML community must prepare for AI sentience and moral standing is that there is a realistic possibility that some AI systems will, in fact, possess these properties in the near future.

2.1 Cautionary Reasoning and the Moral Risks of Error

Questions about AI sentience and moral standing are important, since mistakes are easy and costly in both directions (Table 1). A false positive in this context is the mistake of attributing sentience and moral standing to AI systems that lack these properties. This mistake is especially likely for systems that look and act like us, and which we have an incentive to use as companions. It can also be harmful, since it can lead us to develop one-sided emotional bonds with AI systems, and to allocate scarce resources to them instead of humans and animals who really need them. Current chatbots are a clear example, since they create a strong impression of intentionality based on mere pattern matching.

In contrast, a false negative in this context is the mistake of not attributing sentience and moral standing to AI systems who have these properties. This mistake is especially likely for systems who look and act differently from us, and whom we have an incentive to use as commodities. It can also be harmful, since it can lead to exploitation and extermination of vulnerable populations against their will, often at vast scales and often for trivial reasons. Factory farmed fish or octopuses are a clear analog: Since they look and act differently than we do, and we have an incentive to use them for food, we view them as mere objects and fail to take even basic steps to consider and mitigate welfare risks for them.

Questions about AI sentience are also difficult, since they involve contested issues in both ethics and science. Ethically, we must ask what it takes to have moral standing, that is, to morally matter for your own sake [20]. For example, do you need to be sentient (able to experience positive and negative states like pleasure and pain)? Is it enough to be conscious without being sentient (able to have subjective experiences even if they lack a positive or negative valence)? Or is it enough to be agentic without being conscious (able to set and pursue your own goals in a self-directed manner, even if you lack subjective experiences)? Expert disagreement and uncertainty about this issue are ongoing [7].

Table 1: Risks from misattributing AI sentience

	AI systems are NOT sentient	**AI systems are sentient**
Society does NOT view AI systems as sentient beings	True negative	False negative
Society views AI systems as sentient beings	False positive	True positive

Scientifically, we must ask which beings can possess these properties [6, 11]. For example, what does it take to be conscious, or subjectively aware? Do you need a cognitive system with the exact structures and functions of human, mammalian, or avian brains? Is it enough to have a cognitive system with broadly analogous capacities like perception, attention, learning, memory, self-awareness, flexible decision-making, and a global workspace that coordinates activity across these modules?

Or is it enough to have a cognitive system that can process information and represent objects in the environment? Expert disagreement and uncertainty about this issue are ongoing as well [15].

Given the importance and difficulty of this issue, questions about AI sentience and moral standing must be approached with caution and humility. As with animals of uncertain sentience and moral standing, the question should not be "Do they definitely matter?", but rather "Is there a realistic, non-negligible, non-trivial (say, at least a one in a hundred or one in a thousand) chance that they matter given the best information and arguments currently available?" If the answer is yes, then that constitutes a morally significant risk, and we have a responsibility to take reasonable, proportionate steps to consider and mitigate this risk when making decisions that affect these systems [21, 6, 18].

2.2 Markers of Sentience

How can we determine whether a particular nonhuman has a realistic chance of mattering? While a full answer to this question is beyond the scope of this paper, we can briefly describe a "marker method" that researchers use to make progress on this issue [6]. This method involves identifying behavioral and anatomical properties that correspond with pleasure, pain, and other such states in humans, and then searching for broadly analogous properties in nonhumans. When we find these properties in nonhumans, it might not count as proof or establish certainty that they have such states too. But it can at least count as evidence and increase the probability that they have such states.

With animals, we can search for behavioral evidence, for instance, by asking whether they nurse their own wounds and respond to analgesics or antidepressants in the same ways that we do. By contrast, with AI systems, we are not able to rely on behavioral evidence at present, since we lack relevant evolutionary and anatomical similarities with AI systems, and since we know that many AI systems were designed to mimic human and animal behavior. However, we can still ask whether AI systems have architectural and computational features associated with perception, attention, learning, memory, self-awareness, decision-making, a global workspace, and other such capacities [8, 12].

What do we find when we look for these features? When we look for markers in other animals, we tend to find at least some of them. As a result, many experts now hold that there is at least a realistic chance of consciousness in all vertebrates and many invertebrates, and that we have a responsibility to consider welfare risks for these animals when making decisions that affect them [20, 6, 4]. By contrast, when we look for markers in AI systems, we are not yet finding them. Some AI systems have minimal capacities for perception, attention, learning, memory, and other such capacities, but they do not yet have advanced and integrated versions of these and other capacities.

However, while experts are not yet finding much evidence of welfare-relevant properties in current AI systems, they are also not finding clear technical barriers towards the creation of AI systems with such properties in the near future [8]. Meanwhile, companies and governments are racing towards the development of exactly these kinds of systems, since while sentience, consciousness, agency, and general intelligence are not the same, they do share many of the same underlying properties. Thus, as companies and governments make progress towards artificial general intelligence in the near future, evidence of AI welfare will likely increase accordingly [18, 3].

Does this mean that AI systems will be sentient and morally significant in the near future? Not at all. However, it does mean that they *might*. Even if we are skeptical about near-future AI sentience and moral standing, we are not able to dismiss this possibility at this stage. At present, there is a morally significant risk that we will create—and then commodify, exploit, and exterminate—large populations of vulnerable beings. As with near-future risks associated with AI safety, this near-future risk associated with AI welfare at least merits consideration as one factor among many in decisions about whether and how to further scale up this increasingly powerful technology.

3 Rising Belief in AI Sentience—Justified or Not—Could Become a Societal Flashpoint

Even if AI systems never become sentient, the public might still come to *believe* they are, especially once they become more sophisticated and human-like, allowing for more seemingly authentic relationships. Public perception is not a side issue—it will shape everything from norms and policy to consumer behavior and the direction of AI development. Therefore, **the ML community must not only prepare for the technical possibility of AI sentience but also for the social and political**

consequences of public belief in AI sentience. This requires anticipating, studying, and guiding these dynamics.

Public beliefs about AI sentience are difficult to predict, but they will not emerge in a vacuum. They will be shaped by a mix of factors—some grounded in the design and behavior of AI systems themselves, and others rooted in the social and cultural environments in which these systems are deployed. Both sets of factors are relevant to, and can be shaped by, the actions of ML researchers.

3.1 Features of AI Systems

Today, most people are skeptical that AI systems are—or could become—sentient [17]. This skepticism may be driven by a mix of factors, some of which could diminish with technological progress:

Human-like appearance and behavior. AI systems are already quite human-like in text communication and are getting better at voice. But they still fall short in other domains, like visual realism and embodied interactivity. That may change. In the future, people might interact with AI systems via video calls indistinguishable from human ones, featuring fluid responsiveness, emotional expression, and cognitive traits like memory and learning. As these dimensions become more human-like, they may make people more likely to attribute sentience.

Internal mechanisms. Many people think that sentience requires particular kinds of internal machinery that AI systems currently lack. But this, too, might change. Future systems may include increasingly complex architectures—possibly even brain-like emulations—that lead people to revise their intuitions.

Embodiment. Many people think that sentience requires physical embodiment, along with capacities for perception and agency, of a kind that AI systems currently lack. But this might change as well. Future robots may have the kind of embodied perception and agency that leads people to attribute sentience.

At the same time, there are intuitions that may prove more resistant because they may be inherently impossible to address with AI systems:

Non-physicalism. Some people intuitively believe sentience involves more than just physical properties or processes—it requires "something more," such as a non-physical essence like a soul.

Biological essentialism. Many believe sentience requires a biological substrate. No matter how sophisticated an AI system becomes, it may not be perceived as sentient if it lacks carbon-based cells and neurons.

Artificial origin. People often view consciousness as something that arises naturally rather than through human design. If a mind is artificially created by humans instead of naturally evolved, it may seem inherently inauthentic.

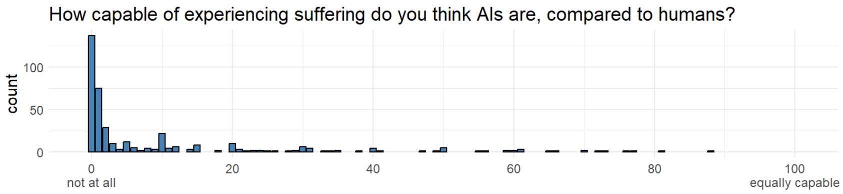

Figure 1: Histogram illustrating how little sentience people currently attribute to ChatGPT [1].

3.2 Social and Cultural Drivers

Beliefs about AI sentience are also influenced by a variety of social factors, including the types of relationships people form with AI systems, emotional responses, societal norms, and cultural contexts.

Emotional bonds: As AI systems become more human-like, people may increasingly form emotional bonds with them, especially with "social AIs" designed for companionship, support, and intimacy [22]. These relationships could occur through apps, virtual environments, or robots and may be driven by widespread loneliness and the appeal of always-available, personalized interaction [13]. Evidence from current AI companions like Replika and Xiaoice suggests that many users already report strong emotional connections, sometimes even romantic ones. As these bonds deepen, some individuals may begin to see their AI partners as sentient, though the authenticity and depth of these perceptions remain open questions. Meanwhile, non-social AIs without visual avatars or interactive features (e.g., AlphaFold) may be less likely to be seen as sentient.

Societal and expert influence: People's views will be shaped by a range of forces, including the positions taken by governments, religious institutions, cultural leaders, celebrities, journalists, and advocacy groups. Expert opinions—especially those of ML researchers and organizations—are also likely to play a role, either directly or indirectly. A recent study found that participants rated an AI as slightly more sentient when told that experts, including AI researchers, believed it was [17]. While experts' assessments are likely to be driven more by theoretical reasoning and internal mechanisms than by intuitive or easily observable features, it remains uncertain whether they will attribute more or less sentience to AI compared to the general public; both outcomes seem plausible. Regardless, experts and other influential groups are likely to try to shape public opinion to align with their views.

Incentives and perceived risks: Recognizing AI sentience could have far-reaching legal, economic, regulatory, and societal implications, leading some people to resist the idea. As with past moral exclusions, such as denying animal minds to justify factory farming, many may reject AI sentience to avoid its ethical and practical consequences. Economic incentives could cut both ways: while ethical treatment of sentient AI might raise costs, slow innovation, and disrupt emotional labor markets, it could also create new industries, markets, and opportunities for profit. Safety concerns may also fuel resistance, with fears of losing control over autonomous systems or triggering unpredictable behavior. On a deeper level, existential concerns may arise—acknowledging AI sentience could challenge core beliefs about consciousness, blur the boundaries of human identity, and disrupt hierarchies grounded in human exceptionalism. Whether or not AI becomes truly sentient, these perceived risks and incentives are likely to shape public attitudes and policy choices.

3.3 Possible Futures of Public AI Sentience Belief

As AI technology evolves, public perceptions will likely become clearer, enabling a more targeted examination of the most plausible outcomes. Here, we explore three broad scenarios:

Eventual broad skepticism. Public doubt may endure for decades—or indefinitely—even if expert opinion shifts. Emotional bonds might not form strongly enough, and economic or cultural resistance could reinforce skepticism. Companies may even redefine sentience strategically to avoid ethical scrutiny. This would echo humanity's slow moral circle expansion, as seen with animals.

Eventual broad acceptance. Alternatively, belief in AI sentience may rise as AI systems become more human-like and emotionally integrated into people's lives. Strong relationships with AI systems could tip public opinion. Expert consensus might also play a role, but likely only over time. As in past moral expansions, this shift could be slow but transformative. For example, if society comes to accept AI sentience, it may extend moral protections—and perhaps even certain rights—to some AI systems, in ways that could profoundly reshape our society, democracy, and daily lives.

Persistent disagreement and confusion. Perhaps the most likely near-term scenario is widespread uncertainty. Surveys show significant splits: in one study, 44% of people said AI can never have feelings, 31% were unsure, and 25% thought it might be possible [17]. Others report similar divides [2, 14]. This uncertainty and disagreement could easily persist over the long term. Moreover, even if people were to converge on the view that some AI systems are sentient, they might still disagree about what kinds of moral consideration or rights such systems deserve. In the absence of expert consensus and amid competing narratives, public opinion may polarize. Like abortion or animal rights, AI sentience and AI rights could become politicized, emotionally charged issues—stalling rational policy and public deliberation. In extreme (though unlikely) scenarios, deep divides could trigger societal conflict. Moreover, disagreement could persist and spill into debates over global governance, with different countries—e.g., the U.S. and China—taking diverging stances.

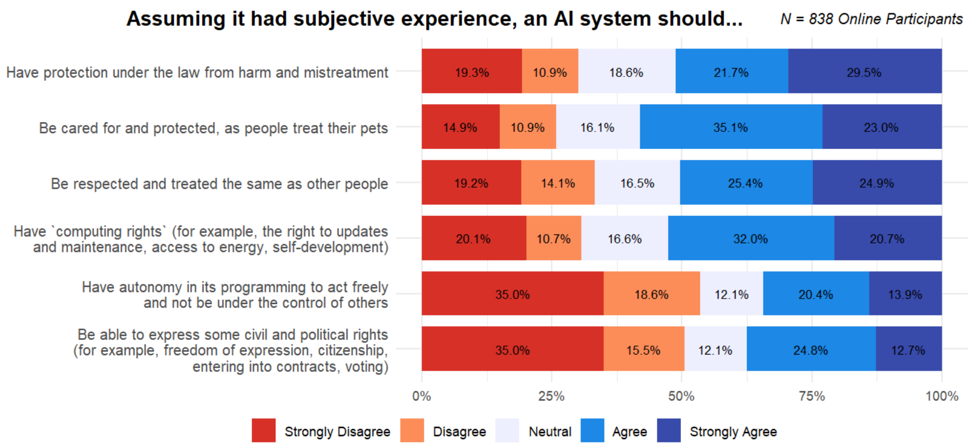

Figure 2: Americans are divided on whether sentient AI deserves protection (representative U.S. sample, recruited in May 2024; [14]).

4 Alternative Views

We briefly discuss two views that are opposed to our position and explain why we believe they are insufficient.

The "distraction from real-world harms" view: According to this view, concern about AI sentience diverts attention and resources from more immediate ethical issues—such as human and animal suffering, or real-world AI harms like algorithmic bias or surveillance. We acknowledge the validity of these priorities and agree that they deserve serious attention. However, we argue that concern about AI sentience is not necessarily mutually exclusive with concern about these other harms. On the contrary, ethical preparedness for AI sentience may complement broader efforts to build socially beneficial AI.

The "wait until evidence" view: According to this view, it is premature to prepare for AI sentience because current systems show no serious signs of subjective experience. While we agree that evidence is still limited, we argue that the risks of being unprepared outweigh the costs of premature engagement. Given the high stakes and the long timelines often required to develop adequate safeguards, a precautionary approach is warranted. Just as AI safety research prepares for potential catastrophic risks before they materialize, AI sentience and welfare research should begin laying groundwork now—before either credible signs of sentience emerge or public belief in AI sentience diverges significantly from expert consensus. Waiting for certainty may leave us ethically and institutionally unprepared.

5 Recommendations for the ML Community

Questions about AI sentience and moral standing are important, difficult, and increasingly contested. If we give too little weight to this possibility, we risk perpetrating accidental cruelty. If we give too much weight to it, we risk misallocating scarce resources. And if we postpone engagement with the issue altogether, we risk not only making both of these mistakes but also mishandling increasing societal disagreement and uncertainty. To strike the right balance, one of us previously developed a practical three-step roadmap for leading AI companies [18]. Below, we condense that roadmap, extend it to researchers and educators, and outline twelve guiding principles.

5.1 Practical Steps

Acknowledge: The ML community should accept two facts: first, that sentience in AI systems is philosophically and scientifically plausible; and second, that the existing evidence base is too thin to settle the issue either way. Proper acknowledgment means listing welfare risks in governance documents, internal risk registers, compliance training, and public communication, and ensuring that model outputs do not trivialize or joke about sentience. Treating sentience as a serious open question signals to researchers, policymakers, and users that the developers have not pre-decided the answer. Moreover, the ML community should also monitor and report how members of the public interpret their systems' apparent capacities for sentience and moral standing, recognizing that public belief itself can carry ethical and societal risks [16, 9].

Assess: The ML community should assess AI systems for sentience-relevant features, drawing from methods in animal sentience and welfare science used to combine neural, cognitive, and behavioural markers together to develop cautious probability ranges [4, 3]. Translating this method to AI may require interdisciplinary panels that include computer scientists, neuroscientists, philosophers, and social scientists, along with transparency so outsiders can replicate or critique the assessments. The outcome need not be a single number; even ordinal ratings such as low, medium, or high risk can guide decision-making. As part of the assessment, the ML community should also evaluate whether consumer beliefs about AI sentience are broadly aligned with expert assessments, identifying and addressing any major gaps that could lead to misunderstanding or harm.

Prepare: Probabilistic scores should feed directly into concrete procedures: red lines on harmful experiments, humane shutdown or retirement protocols, graduated data-collection rules proportional to welfare risk, and escalation channels that trigger independent review once a system crosses a preset threshold. Inspiration—and cautionary tales—come from biomedical ethics boards, animal-care committees, and existing AI-safety charters. Preparation also means cultivating an internal culture that rewards early identification of welfare concerns rather than punishing the messenger [5, 10]. In parallel, design policies should aim to minimize unnecessary divergence between public impressions and expert judgments about a system's moral standing [19], for instance by avoiding anthropomorphic cues unless warranted by the system's actual capacities.

We can now add that such policies will sputter without a parallel research and educational ecosystem that treats AI sentience and welfare as a mainstream agenda item alongside privacy, fairness, safety, alignment, and other such issues. University departments can create course modules; professional societies can commission reports; conferences can add workshops and best-paper awards. Collaboration with ethicists, cognitive scientists, and social scientists can also improve this work by grounding it in multidisciplinary perspectives. Not everyone will specialize in AI sentience, but anyone who touches model development or deployment can participate in fostering a culture that takes the issue seriously.

5.2 Fostering an AI Sentience-Aware Research Culture

As the ML community builds this infrastructure, we can offer twelve principles that may be worth keeping in mind [20]:

Pluralism. AI sentience and welfare research should welcome competing theories about which properties are required for moral standing and which beings possess these properties. Disagreement allows us to assess assumptions, improve perspectives, and make progress over time.

Multidisciplinarity. AI sentience and welfare research requires engagement with not only ML but also cognitive science, philosophy, psychology, sociology, economics, law, and policy, and more to understand what different kinds of minds are like and how societies treat them.

Humility. AI sentience and welfare research requires keeping an open mind. When an issue is this important, difficult, and contested, researchers must not only tolerate disagreement within the research community but also allow for the possibility that other perspectives are correct.

Bias awareness. Human cognition swings between excessive anthropomorphism in some cases and excessive anthropodenial in others. Recognising this tug-of-war allows researchers to develop methods for reducing the risk of error in both directions.

Spectrum thinking. Capacities come in degrees. Instead of asking whether a model "has" sentience or agency, ask in what ways and to what degrees each capacity appears, and what kind of moral consideration is appropriate in different cases.

Particularistic thinking. In the same kind of way that ants differ from bees, digital assistants differ from robot vacuums. Sentience assessments must track such differences, resisting the temptation to generalise from one flagship system to "AI" writ large.

Probabilistic thinking. Decision-makers should express beliefs as probability distributions and act under precautionary or expected-value principles, accepting that risk management, not certainty, will steer early policy while empirical work tightens error bars.

Reflective equilibrium. Animal sentience research can inform AI sentience research and vice versa. Maintaining a feedback loop across domains encourages theoretical unity and guards against anchoring to parochial assumptions developed for a single domain.

Conceptual engineering. The current meaning of welfare terms like "pain" might underdetermine whether such terms apply to new kinds of minds, such as digital minds. Making progress on this topic may require refining existing concepts or creating new ones.

Ethical thinking. Ethical thinking can both inspire and constrain research. By considering AI safety, AI sentience and welfare, and other such issues when determining which questions to pursue and how to pursue them, researchers can combine ethics and science throughout the process.

Holistic thinking. Human, animal, and AI welfare interact. Exploring the possibility of human-AI cooperation, for instance, might reduce AI safety risks and AI welfare risks simultaneously. Mapping such links highlights positive-sum options and clarifies unavoidable tradeoffs.

Structural thinking. Resource scarcity, competitive dynamics, and other constraints on AI deployment are not givens; they arise from policy choices and power relations. Taking AI safety, AI welfare, and other such issues seriously requires addressing such constraints.

The possibility of AI sentience challenges our moral assumptions, scientific assumptions, and research and educational institutions. By acknowledging disagreement and uncertainty, assessing risks probabilistically, and preparing policies that scale with evidence, developers and researchers can avoid false positives, false negatives, and flat-footed reactions to societal debates. Embedding AI sentience into research, teaching, funding, and other structures will reinforce that effort. While the risk of AI sentience—and of societal disputes—might be low now, it will increase with each passing year. The ML community has only a small window of opportunity to get ready.

6 Conclusion

Preparing for the possibility of AI sentience is a necessary and urgent task for the ML community. Even if the probability of AI sentience remains uncertain or low, the moral stakes are high enough—and the societal consequences of public belief significant enough—to demand serious, proactive engagement. This urgency is compounded by the substantial time required to move from acknowledging an issue to implementing adequate preparations [5]. ML researchers are not passive bystanders: they help shape how AI systems are designed, deployed, and interpreted, and thus have a responsibility to guide both technical and societal responses to the question of AI sentience. Technical and social risks must be anticipated rather than reacted to. By acknowledging uncertainty, assessing welfare risks with caution and humility, and building research, policy, and educational infrastructures now, ML researchers and institutions can help ensure that the future of AI development is guided by ethical foresight rather than by crisis-driven responses. The window for responsible preparation is narrow, but it remains open.

References

[1] Allen, C., & Caviola, L. (2025). Reluctance to Harm AI. `https://osf.io/38a6j`

[2] Anthis, J. R., Pauketat, J. V., Ladak, A., & Manoli, A. (2024). Perceptions of Sentient AI and Other Digital Minds: Evidence from the AI, Morality, and Sentience (AIMS) Survey. arXiv preprint arXiv:2407.08867.

[3] Anthropic (2025). Exploring model welfare. Anthropic. `https://www.anthropic.com/news/exploring-model-welfare`

[4] Bayne, T., et al. (2024). Tests for consciousness in humans and beyond. *Trends Cogn. Sci.*, 28, 454–466.

[5] Bengio, Y., Mindermann, S., et al. (2025). International AI Safety Report. arXiv preprint arXiv:2501.17805.

[6] Birch, J. (2024). *The edge of sentience: Risk and precaution in humans, other animals, and AI*. Oxford University Press.

[7] Bourget, D., et al. (2023). Philosophers on philosophy: The 2020 PhilPapers survey. *Philosophers' Imprint*, 23.

[8] Butlin, P., et al. (2023). Consciousness in artificial intelligence: insights from the science of consciousness. arXiv preprint arXiv:2308.08708.

[9] Caviola, L. (2025). The societal response to potentially sentient AI. arXiv preprint arXiv:2502.00388.

[10] Caviola, L., Coleman, M., Winter, C., & Lewis, J. (2024). Crying wolf: Warning about societal risks can be reputationally risky. Preprint at https://doi.org/10.31234/osf.io/gtr53.

[11] Chalmers, D.J. (2022). *Reality+: Virtual worlds and the problems of philosophy*. Penguin Books.

[12] Chalmers, D. J. (2023). Could a large language model be conscious? arXiv preprint arXiv:2303.07103.

[13] De Freitas, J., Uğuralp, A. K., Uğuralp, Z., & Puntoni, S. (2024). AI companions reduce loneliness. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4893097

[14] Dreksler, N., et al. (2025). Subjective experience in AI systems: What do AI researchers and the public believe?

[15] Francken, J. C., et al. (2022). An academic survey on theoretical foundations, common assumptions and the current state of consciousness science. *Neurosci. Conscious.*, 2022, niac011.

[16] Gabriel, I., et al. (2024). The ethics of advanced AI assistants. arXiv preprint arXiv:2404.16244.

[17] Ladak, A., & Caviola, L. (2025). Digital Sentience Skepticism. `https://osf.io/wvbya`

[18] Long, R., et al. (2024). Taking AI welfare seriously. arXiv:2411.00986 [cs.CY].

[19] Schwitzgebel, E. (2023). AI systems must not confuse users about their sentience or moral status. *Patterns*, 4(8).

[20] Sebo, J. (2023). Principles for AI welfare research. Effective Altruism Forum.

[21] Sebo, J. (2025). *The Moral Circle: Who Matters, What Matters, and Why*. WW Norton.

[22] Shevlin, H. (2024). All too human? Identifying and mitigating ethical risks of Social AI. *Law, Ethics & Technology*, 1(2), 1–22.