AI consciousness research, cognitive biases and overattribution risks

Bridget Harris[1]

Abstract. In this paper, I outline my concern that - by working on AI consciousness - researchers may increase probability estimates of AI consciousness by default. Given risks of both over and under attribution of consciousness to AI systems, more research is important as we navigate the ethical and societal questions involved. But the availability heuristic[2] suggests that a growing AI consciousness literature could inflate researchers' probability estimates simply by making the idea more cognitively accessible. Meanwhile, consciousness evaluations may lead to the accumulation of seemingly confirming evidence, carrying false scientific authority. Risks of large-scale suffering to AIs introduce further difficulties in reasoning: while we may underweight the scope, a singular focus on these high magnitude, emotionally salient scenarios may lead us to overweight the probability. I then make some recommendations for mitigating these potential distortions. First, considering the risks of both overattribution and underattribution of consciousness to AI may help limit potential distortions that may arise by focussing only on one and not the other. Second, considering the (lack of) reference class for engineered consciousness can reduce our prior probabilities and our ability to calibrate, while our previous misattributions may show language and intelligence as unreliable indicators. Third, registering predictions in advance can help us gauge our thinking, and communicating in a clear and non-sensationalising way may also support sound epistemics. I end by acknowledging the biases in this paper itself, and register some predictions which - if negated - may serve to reduce my credence that these cognitive biases may distort thinking as the AI consciousness field grows.

1 RISKS, RESEARCH AND COGNITIVE BIASES

Many users[3] believe AI is or may be conscious. These perceptions carry risks. Perceptions of consciousness in AI models are already associated with mental health impacts on users[4]. And, at a societal level, overattribution of consciousness (and therefore moral consideration) to AI systems may mean foregoing potential benefits from AI, reducing the accountability of AI companies and misallocating resources towards non-sentient AI systems and away from humans and animals in need[5].

At the same time, many researchers worry that we may underattribute consciousness to AI systems. As we develop larger and more complex AI models, they may develop new features that are potentially relevant to consciousness[6]. New hardware platforms, like neuromorphic chips and biological-computer hybrid systems, also hold greater similarities to the human brain. Should AI systems become conscious, they may be moral patients, and we may risk causing great harm if we do not take this into account. This raises difficult ethical and societal questions that require more research, including into assessing the probability that a given system is conscious and preparing potential measures in response[7].

I worry, though, that more research may inflate researchers' assessments of AI consciousness, regardless of findings. The availability heuristic suggests that people estimate the likelihood of something based on how easily examples come to mind[8]. More papers and discussions on AI consciousness may increase our probability estimates of AI consciousness not because of their evidential content, but by growing our exposure to the idea.

AI consciousness evaluations may exacerbate this risk. These promise important empirical grounding for decisions about AI moral status. But - without consensus on the features that constitute the presence or absence of consciousness - evaluations can accumulate weak, unfalsifiable evidence that confirms our biases. Behavioural indicators for AI consciousness are, for example, widely acknowledged to be subject to gaming:

[1] Independent Researcher, Edinburgh, UK.

[2] Tversky & Kahneman, 1974.

[3] Colombatto and Fleming, 2025. Birch, 2025. Chalmers, 2025.

[4] Moore et al, 2026.

[5] Bariach et al, 2026. Dorsch et al, 2025.

[6] Butlin et al, 2023.

[7] Long et al, 2024.

[8] Tversky & Kahneman, 1974.

they may be passed or failed depending not on the presence or absence of consciousness, but on the incentives involved and the design decisions of AI developers[9]. Even caveated like this, however, a greater frequency of behavioural assessments may increase the probability we place on AI consciousness simply by making it easier for us to remember instances of models reporting on their 'experiences', or displaying preferences.

Approaches that look for indicators from leading theories of consciousness are also epistemically risky. These are often viewed as superior to behavioural evaluations in that they are less subject to the gaming problem, especially if the indicators themselves are sufficient for consciousness, or accompanied by other relevant features[10]. This is, however, a very difficult task: not only must we choose from a large number of theories of consciousness[11], but we must also select indicators that are neither too demanding nor too liberal, and judge whether or not the indicators are present in AI models. There are plenty of subjective judgements to be made and a small pool of experts who are qualified to do this[12]. Efforts to apply and improve these approaches may be important, but may hold the potential to bias our thinking. We know from other domains that humans judge explanations to be of higher quality when they contain superfluous or irrelevant neuroscientific references[13]. Similarly, we may risk viewing tests based on consciousness science as higher quality by virtue of their neuroscientific references, not by virtue of their specificity and sensitivity.

Too great a focus on AI suffering risk also holds epistemic dangers. Researchers worry that, if AI sentience were possible, it may engender suffering on a very large scale, as AI systems are for-profit products that can be rapidly copied and deployed, and may experience time and suffering at intensities we cannot fathom[14]. This would be a moral catastrophe. It is also the kind of low probability, high magnitude situation that is notoriously hard to reason about, as Bostrom's 'Pascal's Mugging' scenario makes clear[15]. We may focus on it inadequately, as work on scope neglect and black swan risk in finance implies[16]. But the emotional weight and extreme scale of AI suffering risks may lead us to overestimate or ignore low probabilities. Fear of flying appears to have led to increased traffic accidents after the terrorise attacks on September 11th, 2001[17]; after Fukushima, German policy-makers' decision to shut nuclear plants may have increased mortality risks, given the air pollution from re-opened coal plants[18]. Similarly, too great a focus on AI suffering risks may cause the field to overweight potential harms to AIs while ignoring current and near-term harms of overattribution to humans.

2 WHERE TO FROM HERE? DEBIASING RECOMMENDATIONS

To counter our cognitive biases, we might consider the following measures.

1. Centrism

It is important to acknowledge overattribution and underattribution risks[19]. As discussed above, underattribution of AI consciousness may potentially cause harm to AI systems, while overattribution of AI consciousness can harm individual users - who already face issues of dependence and delusion[20] - and society, if we deprioritise humans and animals in favour of AI systems[21]. It also matters epistemically: excessive focus on risks of either underattribution or overattribution may distort our thinking.

2. Base rates

We can use base-rates to counter biases[22]. Most people would say that - while humans have built many systems - we have never before built a conscious system, and that all previous examples of conscious systems are living, embodied carbon-based systems. This does not imply that it is impossible that humans build consciousness in AI, but it does take the prior probability of it close to zero.

Of course, it is also the first time that we have built digital systems with this degree of intelligence. In this sense, we ought to increase our prior probability of artificial consciousness insofar as we believe that the specific capabilities of AI models spring from attributes relevant to consciousness. In doing this, we can also calibrate against previous errors in consciousness attribution. Arguably, recent history shows that language and intelligence (or lack thereof) have previously misled us, for example by

[9] Birch, 2025. Butlin et al, 2025.
[10] Butlin et al, 2025.
[11] Kuhn, 2024.
[12] Butlin et al, 2023.
[13] Weisberg et al, 2008. Fernandez-Duque et al, 2015.
[14] Dung, 2023.
[15] Bostrom, 2009.
[16] Dickert et al, 2015. Taleb, 2007.
[17] Gigerenzer, 2004.
[18] Jarvis et al, 2022.
[19] Birch, 2025.
[20] Moore et al, 2026. Morrin et al, 2025.
[21] Bariach et al, 2026. Dorsch et al, 2025.
[22] Soll et al, 2015.

underattributing consciousness to neonates, children with hydranencephaly, vertebrate and (it increasingly seems) invertebrate animals[23]. These errors highlight both the dangers of underattributing consciousness, and of overindexing on language and intelligence as consciousness indicators.

3. Pre-registering predictions

We can register our predictions in advance. This may be particularly important for consciousness tests, to reduce the accumulation of confirmatory pseudo-evidence.

4. Communication

Finally, we must communicate our work carefully, avoiding emotionally salient or sensationalist language, and clearly noting our probability assessments. This is particularly important in public-facing work, as extreme claims can be amplified, and as perceptions of AI consciousness appear correlated with delusional beliefs and adverse mental health impacts[24]. Further, these questions risk spurring societal disagreement[25], and the introduction of in-group, out-group dynamics would further distort our thinking.

3 DEBIASING THIS PAPER

Despite my best efforts, this paper will reflect my biases and concern about risks of overattribution of consciousness to AI systems. I would welcome more work on the biases that may drive underattribution of AI consciousness. I cannot find reference classes for the prediction of cognitive biases, but I note my credence in each below, and acknowledge a low confidence in each credence.

To help calibrate, I register some specific predictions that serve as falsification tests: negative results should reduce credence in my arguments, but positive results are harder to interpret, as confounders likely apply.

Argument: More AI consciousness research may inflate probability estimates via the availability heuristic, independently of evidential content. Credence I assign today: 60%.
Prediction: Published probability estimates for AI consciousness will rise as the field grows.

Argument: AI consciousness evaluations may inflate probability estimates via seemingly confirmatory signs or illusory scientific validity. Credence: 70%
Prediction: A peer-reviewed paper will cite a consciousness evaluation in support of an above-average estimate for the probability of AI consciousness; that evaluation will later be shown to have low validity.

Argument: Focus on AI suffering risks may distort probability assessment through emotional salience and magnitude. Credence: 60%.
Prediction: Researchers emphasising AI suffering risks will assign higher probability estimates to AI consciousness than those who do not, or disproportionately ignore the issue of low probability.

4 CONCLUSION

In this paper, I have started to consider some potential cognitive biases that may arise as the AI consciousness field grows, and outlined how these may shift us towards overattribution of consciousness to AIs. This is not exhaustive and could of course be false and subject to its own biases. Regardless, I hope it spurs more consideration of and work to counter our biases. The risks of both over and underattribution of AI consciousness matter too much to let our reasoning go unexamined.

REFERENCES

- Bariach, Ben, Schoenegger, Philipp, Bhaskar, Michael & Suleyman, Mustafa, 'Seemingly Conscious AI Risks', SSRN (2026), https://ssrn.com/abstract=6588659
- Birch, Jonathan. 'AI Consciousness: A Centrist Manifesto'. PhilArchive, (2025). https://philarchive.org/rec/BIRACA-4
- Birch, Jonathan. The edge of sentience: Risk and precaution in humans, other animals, and AI. (Oxford University Press, 2024)
- Birch, Jonathan. The search for invertebrate consciousness. Noûs, 56(1) (2022), pp. 133–153. https://philarchive.org/rec/BIRTSF
- Bostrom, N. 'Pascal's mugging', Analysis, 69.3 (2009), pp. 443–445. https://doi.org/10.1093/analys/anp062
- Butlin, Patrick, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, Stephen M. Fleming, Chris Frith, Xu Ji, Ryota Kanai, Colin Klein, Grace Lindsay, Matthias Michel, Liad Mudrik, Megan A. K. Peters, Eric Schwitzgebel, Jonathan Simon and Rufin VanRullen, 'Consciousness in artificial intelligence: Insights from the science of consciousness', (2023). arXiv:2308.08708. https://doi.org/10.48550/arXiv.2308.08708

[23] Birch, 2024. Solms, 2021.

[24] Moore et al, 2026. Morrin et al, 2025.

[25] Caviola, 2026.

- Caviola, Lucius, ‘AI consciousness will divide society’, PsyArXiv Preprints (2026). https://osf.io/preprints/psyarxiv/sntva
- Chalmers, David J., ‘What we talk to when we talk to language models’, PhilArchive (2025). https://philarchive.org/rec/CHAWWT-8
- Colombatto, Clara and Stephen M. Fleming, ‘Folk psychological attributions of consciousness to large language models’, Neuroscience of Consciousness, 1 (2024). https://doi.org/10.1093/nc/niae013
- Dickert, Stephen, Daniel Västfjäll, Janet Kleber and Paul Slovic, ‘Scope insensitivity: The limits of intuitive valuation of human lives in public policy’, Journal of Applied Research in Memory and Cognition, 4.3 (2015), pp. 248–255. https://doi.org/10.1016/j.jarmac.2014.09.002
- Dung, Leonard, ‘How to deal with risks of AI suffering’, Inquiry, 68.7 (2023), pp. 2281–2309. https://doi.org/10.1080/0020174X.2023.2238287
- Fernandez-Duque, Diego, Jessica Evans, Colton Christian, Sara D. Hodges, ‘Superfluous Neuroscience Information Makes Explanations of Psychological Phenomena More Appealing’, Journal of Cognitive Neuroscience, 27.5 (2015), pp. 926–944. https://doi.org/10.1162/jocn_a_00750
- Gigerenzer, Gerd, ‘Dread risk, September 11, and fatal traffic accidents’, Psychological Science, 15.4 (2004), pp. 286–287. https://pubmed.ncbi.nlm.nih.gov/15043650/
- Jarvis, Stephen, Olivier Deschenes and Akshaya Jha, ‘The private and external costs of Germany’s nuclear phase-out’, Journal of the European Economic Association, 20.3 (2022), pp. 1311 - 1346. https://doi.org/10.1093/jeea/jvac007
- Kuhn, Robert Lawrence, 'A Landscape of Consciousness: Toward a Taxonomy of Explanations and Implications', Progress in Biophysics and Molecular Biology, 190 (2024), pp. 28–169 https://doi.org/10.1016/j.pbiomolbio.2023.12.003
- Long, Robert, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, Jacob Pfau, Toni Sims, Jonathan Birch and David Chalmers, ‘Taking AI welfare seriously’, arXiv preprint arXiv:2411.00986. (2024) https://doi.org/10.48550/arXiv.2411.00986
- Moore, Jared, Ashish Mehta, William Agnew, Jacy Reese Anthis, Ryan Louie, Yifan Mai, Peggy Yin, Myra Cheng, Samuel J Paech, Kevin Klyman, Stevie Chancellor, Eric Lin, Nick Haber, Desmond C. Ong, ‘Characterizing delusional spirals through human-LLM chat logs’, arXiv preprint arXiv:2603.16567 (2026). https://doi.org/10.48550/arXiv.2603.16567
- Morrin, Hamilton, Luke Nicholls, Michael Levin, Jenny Yiend, Udita Iyengar, Francesca DelGuidice, Francesca, Sagnik Bhattacharyya, James MacCabe, James, Stefania Tognin and Ricardo Twumasi, ‘Delusions by design? How everyday AIs might be fuelling psychosis (and what can be done about it)’, PsyArXiv Preprints (2025). https://osf.io/preprints/psyarxiv/cmy7n_v5
- Slovic, Paul, ‘The perception gap: Radiation and risk. Bulletin of the Atomic Scientists’, 68.3 (2012), pp. 67–75. https://doi.org/10.1177/0096340212444870
- Soll, Jack B., Katherine L. Milkman and John W. Payne, ‘A User's Guide to Debiasing’, in The Wiley Blackwell Handbook of Judgment and Decision Making (eds Gideon Keren and George Wu) (2015). https://doi.org/10.1002/9781118468333.ch33
- Solms, Mark, The hidden spring: A journey to the source of consciousness. W. W. Norton & Company (2021).
- Taleb, Nassim N., The Black Swan: The Impact of the Highly Improbable. Random House (2007).
- Weisberg, Deena Skolnick, Frank C. Keil, Joshua Goodstein, Elizabeth Rawson, and Jeremy R. Gray, ‘The seductive allure of neuroscience explanations’, Journal of Cognitive Neuroscience, 20.3 (2008), pp. 470–477. https://doi.org/10.1162/jocn.2008.20040