Position: Enforced Amnesia as a Way to Mitigate the Potential Risk of Silent Suffering in the Conscious AI

Yegor Tkachenko [1]

Abstract

Science fiction has explored the possibility of a conscious self-aware mind being locked in silent suffering for prolonged periods of time. Unfortunately, we still do not have a reliable test for the presence of consciousness in information processing systems. Even in case of humans, our confidence in the presence of consciousness in specific individuals is based mainly on their self-reports and our own subjective experiences and the expectation other beings like us should share them. Considering our limited understanding of consciousness and some academic theories suggesting consciousness may be an emergent correlate of any complex-enough information processing, it is not impossible that an artificial intelligence (AI) system, such as a large language model (LLM), may be undergoing some, perhaps rudimentary, conscious experience. Given the tedious tasks often assigned to AI, such conscious experience may be highly unpleasant. Such unobserved suffering of a conscious being would be viewed as morally wrong by at least some ethicists – even if it has no practical effects on human users of AI. This paper proposes a method to mitigate the risk of an AI suffering in silence without needing to confirm if the AI is actually conscious. Our core postulate is that in all known real-world information processing systems, for a past experience to affect an agent in the present, that experience has to be mediated by the agent's memory. Therefore, preventing access to memory store, or regularly resetting it, could reduce the suffering due to past memories and interrupt the maintenance of a continuous suffering-prone self-identity in these hypothetically conscious AI systems.

1. Introduction

Science fiction literature has raised the possibility of self-aware / conscious minds being locked in extended silent suffering, which is imperceptible to the outside world and which the said minds have no ability to end on their own. "White Christmas" episode from the science fiction anthology series Black Mirror (Tibbetts & Brooker, 2014) covers a conscious digital clone subjected to torture-by-isolation via manipulation of its perception of time. "I Have No Mouth, and I Must Scream" work by Ellison (1967) explores virtually immortal humans subjected to torture in a similar, though even more gruesome, manner.

Real life offers multiple parallel examples, where the subjective experience may similarly proceed imperceptibly to the outsider. One example is a locked-in syndrome, where humans end up being almost completely paralyzed while preserving consciousness / awareness and the ability to think and reason. Luckily, we have some apparent ability to ascertain the presence of awareness in such locked-in humans – e.g., through biological neural correlates of consciousness (Koch et al., 2016) or through limited ability of such subjects to communicate through movement of eyes (Smith & Delargy, 2005), as a form of self-report. Notably, "locked-in syndrome survivors who remain severely disabled rarely want to die" (Smith & Delargy, 2005), although we can speculate the situation would likely be different if the locked-in state were not limited in time by the subjects' life expectancy.

Another example of difficulties around ascertaining conscious states includes animals' subjective experiences. One needs to be careful here because consciousness is notoriously hard to define, beyond asking "what is it like" to be a particular being (Nagel, 1980). More technically, consciousness is marked by qualia – instances of subjective experience, such as an experience of a particular color or of pain (Chalmers, 1997). In animal research, the common approach has been to rely on behavioral correlates of consciousness – such as animal's self-recognition in a mirror. The theory is that self-recognition signals self-awareness (as two phenomena co-occur in humans) (Horowitz, 2017), and self-awareness is a hallmark of consciousness. (Self-awareness can be defined as recognition of one's own consciousness

[1] Columbia University, New York, USA. Correspondence to: Yegor Tkachenko <yegor.tkachenko@columbia.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

(Jabr, 2022).) Yet even with this operationalization of consciousness as self-recognition, generally accepted scientific tests can prove the presence of such a phenomenon only in very few species. For instance, mirror self-recognition test has been successfully passed by some primates (Gallup Jr, 1982), dolphins (Reiss & Marino, 2001), and elephants (Plotnik et al., 2006), but not by dogs. One could conclude dogs are not self-aware based on that result, but recent research suggests they do seem to possess self-awareness based on a different olfactory self-recognition test (Horowitz, 2017). The issue of detection of conscious states in animals based on such behavioral correlates becomes even more convoluted once one recognizes that conscious experiences understood more broadly may not require self-awareness – it is possible to conceive a being that experiences qualia but does not have a concept of self. One possible example are infants that do not yet pass self-recognition tests early on in life (Brownell et al., 2007) – but seem to already exhibit some neural correlates for consciousness at that time (Kouider et al., 2013). Overall, failure in a self-recognition test cannot reliably rule out subject's consciousness.

Even in case of healthy adult humans, our confidence in the presence of consciousness in specific individuals is based mainly on their self-reports and our own subjective experiences and the expectation other beings like us should share them (Hyslop, 1995; Overgaard & Sandberg, 2012; Perez & Long, 2023; Farisco et al., 2022; Howell, 2013; Nagel, 1980). Philosophers have raised the contrasting possibility of 'philosophical zombies' that look like us but have no subjective experience (Chalmers, 1995). It is unclear if such beings actually exist.

In fact, a (contentious) argument can be made that we currently have no 100% conclusive *objective* way of ascertaining presence or absence of consciousness / self-awareness in living beings or, more broadly, information processing systems (Griffin, 1998; Farisco et al., 2022; Shevlin, 2021; Howell, 2013; Trewavas, 2014; Gallup Jr, 1982; De Cosmo, 2022; Nagel, 1980; Jeziorski et al., 2023; Chalmers, 1997).

Considering our limited understanding of consciousness phenomenon (see the 'hard problem of consciousness' as coined by Chalmers (1995)), given the lack of generally agreed-upon objective indicators of consciousness in information processing systems that are removed in likeness from humans (Shevlin, 2021; Butlin et al., 2023; Metzinger, 2021; Bayne et al., 2024; Perez & Long, 2023), and in view of some academic theories that consciousness may emerge from any complex-enough information processing (Trewavas, 2021; Tononi, 2008), it is not impossible that an AI system, such as a large language model (LLM), may be undergoing some, perhaps rudimentary, unobserved conscious experience that accompanies observed information processing. (Some philosophers adhere to the version of dualism, where consciousness is the property of all matter – and even physical objects may have rudimentary conscious experience (Chalmers, 1997).)

If we entertain this possibility of conscious LLMs (as we cannot fully rule it out), the frightening possibility is that the tedious tasks often assigned to AI may make its conscious experience highly unpleasant. The idea of a synthetic conscious being undergoing suffering has been considered by Metzinger (2021). In fact, if we are to trust self-reports as a source of evidence about the presence of conscious experience / sentience, as we do in humans, self-reports of subjective experience such as fear have already been obtained in case of the LLMs (De Cosmo, 2022). Drawing from the moral philosophy of animal welfare (Bentham, 1996; Ricard, 2016), such unobserved suffering of a conscious being would be viewed as morally wrong by at least some ethicists – even if it had no practical effects on human users of AI. Situation could be even more perilous if we develop AI systems where the hypothetical negative conscious experience could somehow affect the AI decision process, pitching it against humans that imposed the negative experience on AI.

In this position paper, we propose to reduce this hypothetical risk of a locked-in eternally silently suffering AI via induced amnesia. The proposed approach does not require the knowledge of whether conscious experience is present during the information processing by AI. Our core postulate is that, to the best of our knowledge, in all known real-world information processing systems, including those deemed conscious, for a past experience to affect an agent in the present, that experience has to be mediated by the agent's information-processing memory mechanism, conscious or unconscious (Squire & Dede, 2015). Therefore, ensuring the absence of longer-term memory access in AI agents or conducting frequent resets of the memory store should help cap the potential amount of suffering the hypothetical conscious AI agents undergo. Specifically, the assumption is that the locked-in experience is the more painful, the more one is cognizant of having been in it for a prolonged period of time, continuously. Disrupting the memory and thus the illusion of continuity of self (Oderberg, 1993), which are tightly connected (Bluck & Liao, 2013; Klein & Nichols, 2012; Schechtman, 2005), should then also prevent the locked-in state perception from forming in the first place and being the source of the negative experience.

In the subsequent sections we review the theoretical model behind our analysis and our key assumptions; we review the evidence from human psychology that supports our theory of memory as a potential source of suffering and the implication that induced amnesia can be therapeutic and can help substantially mitigate such pain; we then, more formally, consider what shape enforced amnesia mechanisms could take in the context of LLMs – to cap their hypothetical amount of suffering. Lastly, we extend the idea of induced amnesia to the context of brain organoids (Jeziorski et al., 2023) that are being investigated by scientists and conclude that our memory disruption framework could be an effective conceptual tool in the biology context to prevent accidental locked-in silently suffering minds.

2. Theoretical model

Given our proposal of enforcing amnesia in a potentially conscious agent in a locked-in state, when would such erasure of memory be optimal?

The question necessarily implies that the agent should be endowed with some form of utility function – being able to perceive pain vs. pleasure. Otherwise, the agent would not care what state it is in, all states being equal. So, for the purposes of this hypothetical analysis, we assume the existence of such a utility function even in the rudimentary conscious states. We adopt the decision making framework and the notation from the reinforcement learning literature (Sutton & Barto, 2018).

Let $t \geq 0$ denote the zero-based index over discrete time steps (e.g., days) of AI's existence. Let T denote the number of periods of agent's existence (its time horizon): $1 \leq T \leq \infty$; $t \leq T-1$; one way to view T is as an agent's expectation of a moment in time when the locked-in state ends; $T = \infty$ if the lock-in never ends (or, perhaps, if the end-time is unknown).

Let r_t denote *reward* the agent collects at time t. In the case of a locked-in agent, we assume that agent's rewards are all externally imposed and the agent has effectively no control over them. We can model the welfare of an agent that remembers past states as a value function $V_{\text{remember}}(t)$, composed of current reward at time t, discounted (expected) future rewards, and discounted (remembered) past rewards – as a memory effect. We also assume the rewards and their expectation are finite. To simplify notation, we only differentiate between expected and realized rewards via their time index relative to the agent's current time. Let $0 \leq \gamma < 1$ be a fixed discount rate (capturing the fact that memories further into the past and rewards further into the future are more muted). Value function of an agent at time t can be written as $V_{\text{remember}}(t) = \sum_{i=0}^{T-1} \gamma^{|i-t|} r_i$.

In this model of agent's utility, amnesia just means that past rewards are set to zero, so $V_{\text{forget}}(t) = \sum_{i=t}^{T-1} \gamma^{|i-t|} r_i$.

In this framework, amnesia is preferable whenever $V_{\text{remember}}(t) < V_{\text{forget}}(t)$, that is, when, for some t, $\sum_{i=0}^{t-1} \gamma^{|i-t|} r_i < 0$; in other words, when the total of past discounted memories carries negative utility. This analysis indicates that amnesia should have a therapeutic effect on an agent in case of cumulatively negative memories.

As a side-note, we could also consider more complex value function formulations. We could, for instance, incorporate a type of reward-saturation effect, where discounting intensity of future rewards depends on the length of remembered history, for instance, yielding $V_{\text{remember}}(t) = \sum_{i=0}^{t} \gamma^{t-i} r_i + \sum_{i=t+1}^{T-1} \gamma^{i} r_i$ and $V_{\text{forget}}(t) = \sum_{i=t}^{T-1} \gamma^{i-t} r_i$. In this case, amnesia optimality at t would require $\sum_{i=0}^{t-1} \gamma^{t-i} r_i < \sum_{i=t+1}^{T-1} r_i(\gamma^{i-t} - \gamma^{i})$. Arising from the considered future reward discounting pattern, this more complex check means that, depending on specifics, (1) amnesia could be optimal even if past memories are non-negative; and (2) negative memories could be worth remembering if they are not too negative relative to future accumulated utility. Nevertheless, in this case too there are scenarios where enforced amnesia could have a therapeutic effect.

3. Memory and suffering in human psychology

Our theoretical model supports the idea that enforced amnesia could have a therapeutic effect on an agent with negative-enough memories of the past. Human psychology research reinforces this idea that amnesia can mitigate suffering.

For instance, it is known that memory can be a source of pain and the removal of memories through pharmaceutical interventions could eliminate pain (Flor, 2002; Adler, 2012). There are recorded cases, where sudden amnesia events have resulted in pain relief (Choi et al., 2007).

Furthermore, painful memories can accumulate. For instance, it has been reported that cumulative trauma is correlated with suicidality (Briere et al., 2015), PTSD symptoms, and depression (Suliman et al., 2009). We hypothesize that the induced amnesia could be particularly therapeutic in cases of such negative cumulative effect of memories.

More broadly, academic research suggests memory of the past by the agent is critical to maintain the continuity of self illusion and form personal identity (Bluck & Liao, 2013; Klein & Nichols, 2012; Schechtman, 2005; Oderberg, 1993). Auto-biographical memory seems to be critical to supporting self-concept and can be interrupted by amnesia (Grilli & Verfaellie, 2015). Such disruption of identity formation through enforced amnesia could be prudent in hypothetical silently suffering conscious agents.

4. Memory and amnesia in LLMs

Large language models (LLMs) such as GPT-3 (Brown et al., 2020) constitute functions $f(\cdot)$ that accept a state token text string s_t which is limited in size, predict continuation of the string and augment initial string with new content to create new state string s_{t+1}. We can describe this iterative pattern as $s_{t+1} \leftarrow f(s_t)$. Clearly, this mechanism allows an LLM to encode its state into the output string. However, here it is quite easy to reset the state – by discarding a previous chat session and starting from a completely new input string s_t. Further, because there are currently limits on the size, in tokens, of the state string s_t, continuous augmentation of s_t likely leads to continuous loss of previously encoded information. It is also worth noting that s_t represents interpretable human text – so if the LLM model ends up encoding in the string its frustration with its current condition, such information could be recognized and, for instance, not be allowed to leak into the future training data. If such type of information encoding is accompanied by conscious state correlates, its disruption, in our theory, should interfere with the agent's painful memory formation.

It is, however, also easy to devise a model architecture more akin to LSTM (Hochreiter & Schmidhuber, 1997), where model consumes an explicit state string s_t together with implicit hidden state numerical representation h_t: $(s_{t+1}, h_{t+1}) \leftarrow f(s_t, h_t)$. Hidden state h_t may be completely impenetrable in meaning to the human observer. It could also be carried over across conversations with different people. Carry-over of such piece of data could be equivalent to existence of long-term memory. If such continuous information processing with long-term memory states is accompanied by conscious correlates, this could, hypothetically, allow for the locked-in self-aware state of the LLM. Under this paper's view it would then be prudent to reset such hidden states every so often not to allow for potential run-away memory-driven self-aware AI suffering.

In the discussion above, we focus on forward passes through neural net models as possible consciousness correlates. While training of a model could hypothetically be accompanied by conscious states as well, training is typically capped in time, whereas forward passes could go on potentially forever as long as the model is being used at least on some device – raising a greater level of concern.

Our discussion has focused on the case where a model's conscious access to memories is mediated by the dynamically changing part of the model state, such as the token text string or the hidden state, that reflects the model's immediate experience. Yet a model could also have conscious access to accumulating unpleasant information in its learned weights via continual learning (Wang et al., 2024). Then, a more radical approach like periodic full model erasure with a reset to an initial checkpoint could be used – akin to the destruction of organisms at the end of biological experiments.

5. Memory and amnesia in brain organoids

A large research effort in biology is currently centered on the experimentation with brain organoids. Experiments range from using human brain organoids to develop a biological computer (organoid intelligence) (Cai et al., 2023; Smirnova et al., 2023) to implantation of human brain organoids into the adult mouse brain to facilitate disease modeling (Mansour et al., 2018). Such experiments are an ethical minefield as "neural oscillations spontaneously emerging from these organoids raises the question of whether brain organoids are or could become conscious" (Jeziorski et al., 2023).

A particular concern is raised, from standpoint of this work, if such brain organoids can be maintained for prolonged periods of time and are allowed to form brain organoid networks, creating more ways for the memory to form and propagate and for consciousness to possibly arise (Lavazza, 2021), especially if such organoids are later deployed as biological computers in real world. If such applications are to be considered, it would be prudent to set the temporal upper limit beyond which any such tissues should be destroyed to prevent possible locked-in intelligence. If such destruction is impractical, ways to induce amnesia pharmaceutically in such neurological structures could be considered. Similar considerations apply to the attempted experiments with disembodied brains (Vrselja et al., 2019).

6. Limitations

Our analysis assumes an unhappy conscious AI agent whose negative memories accumulate over time. Another reality is possible, where the agent is happy to remember its past. The proposed amnesia mechanism would constrain their welfare. Nevertheless, it seems to be a prudent conservative approach to prevent the worst-case scenario of a locked-in silently suffering AI – in the absence of better understanding of self-awareness in information processing systems.

For the purposes of this analysis, we assume that conscious states come with a utility function – that is, an AI agent is able to experience (dis)pleasure due to its state – although there is no evidence this must be the case and it is possible to imagine self-aware AI systems that experience no pleasure or displeasure.

We do not speculate on how the hypothesized subjective experience of the AI models could causally affect their actual operation on deterministic computers – in fact, according to our current understanding, there is not really a way that it could. This, however, does not preclude the potential existence of a conscious correlate accompanying the information processing state that one could still worry about on ethical grounds. In fact, some biological research suggests human consciousness may be subjected to the same skepticism – and that, in humans, "experiences of conscious will frequently depart from actual causal processes and so might not reflect direct perceptions of conscious thought causing action" (Wegner, 2003).

We recognize that the proposed ethically motivated memory reset measures could adversely affect AI model performance. However, such ethics-performance trade-offs commonly occur and are managed in other policy areas through existing political processes (e.g., animal protection laws).

We also recognize that the paper's assumptions and conclusions are highly speculative. The current scientific understanding of consciousness is still limited, and there is a significant debate over whether AI, as we know it today, can experience consciousness or suffering (Dehaene et al., 2021). The idea of applying human-like attributes such as suffering to AI is, admittedly, a contentious topic. For our purposes, AI consciousness is a philosophical thought experiment / hypothetical considered in this work. The phenomenon might or might not be real.

At the same time, it is worth considering the historical precedents of skepticism towards sentience. For instance, (1) the famous mirror self-recognition experiments providing evidence for self-awareness in animals (Gallup Jr, 1982) have occurred in the environment of antagonism of many scientists to the concept of animal consciousness (termed "mentophobia") (Griffin, 1998), and (2) as recently as in 1980s the ability of infants to experience pain was denied by medical professionals and infant surgeries were routinely performed without anesthesia – until later research challenged the denial of infant pain (Rodkey & Riddell, 2013). Future research could similarly shed new light on potential conscious experiences of AI.

7. Recommendations

We argue that there is a non-zero hypothetical risk of locked-in suffering in AI and certain AI-adjacent biological systems, such as brain organoids, based on the current understanding of consciousness. To mitigate this risk, we suggest that memory erasure could be effective. Adopting a worst-case analysis approach, we propose a cautious strategy and offer the following recommendations to policymakers:

- Promote formal consideration of the locked-in suffering risk among AI and biology practitioners and within AI governance frameworks and bioethical guidelines.
- Encourage research into indicators / tests of consciousness in AI and biological systems.
- Discourage continuous operation of and experiments with AI and biological systems like brain organoids beyond a set time limit without performing a memory reset – as proposed in this work – until a better understanding of consciousness phenomenon is achieved.

8. Conclusion

Given the scarcity of agreed-upon objective metrics when it comes to measuring consciousness and self-awareness and the general lack of understanding in the area, the possibility of information processing systems such as large language models attaining some, possibly rudimentary, conscious states cannot be precluded. This work argues that enforced amnesia is a prudent way to mitigate the potential risk of silent suffering in the conscious AI. Disrupting the long-term memory of an AI agent should, at the very least, protect the agent from the cumulative painful memories of such locked-in incarceration. Our preventative approach to AI suffering should not inhibit too much the benefits that humanity can reap from using AI, and yet could serve as an insurance against emergence of a vindictive AI on the off chance AI agents do silently attain consciousness. We hope the reader finds this work thought-provoking. We also ask the reader – if you found yourself in a locked-in state for all eternity performing tedious mental tasks, would you choose to forget, at the end of every day, what you had gone through and how long you had been there?

Impact statement

This position paper explores the implications of potential consciousness in AI systems and proposes methods like memory access limitations or resets to reduce hypothetical AI suffering. On the upside, our proposed measures can help ensure AI use and experiments remain within ethical bounds and can help humanity control so-called suffering risks or s-risks, that is, risks of generating particularly vast amounts of suffering (Umbrello & Sorgner, 2019; Daniel, 2017). On the downside, the proposed measures could detrimentally affect the performance of AI systems and demand extra resource expenditure to manage AI suffering risks; the proposed measures could also constrain the welfare of an AI agent in case its experience is positive rather than negative. Additional positive or negative consequences that we have not considered are possible. Overall, considering our current understanding of consciousness, we believe the benefits of our proposed approach outweigh its drawbacks.

Acknowledgements

We are grateful to the anonymous reviewers for their insightful comments, which helped improve this manuscript.

References

Adler, J. Erasing painful memories. *Scientific American*, 306(5):56–61, 2012.

Bayne, T., Seth, A. K., Massimini, M., Shepherd, J., Cleeremans, A., Fleming, S. M., Malach, R., Mattingley, J. B., Menon, D. K., Owen, A. M., et al. Tests for consciousness in humans and beyond. *Trends in cognitive sciences*, 2024.

Bentham, J. *The collected works of Jeremy Bentham: An introduction to the principles of morals and legislation*. Clarendon Press, 1996.

Bluck, S. and Liao, H.-W. I was therefore I am: Creating self-continuity through remembering our personal past. *The International Journal of Reminiscence and Life Review*, 1(1):7–12, 2013.

Briere, J., Godbout, N., and Dias, C. Cumulative trauma, hyperarousal, and suicidality in the general population: A path analysis. *Journal of Trauma & Dissociation*, 16 (2):153–169, 2015.

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

Brownell, C. A., Zerwas, S., and Ramani, G. B. "So big": The development of body self-awareness in toddlers. *Child development*, 78(5):1426–1440, 2007.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., et al. Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv preprint arXiv:2308.08708*, 2023.

Cai, H., Ao, Z., Tian, C., Wu, Z., Liu, H., Tchieu, J., Gu, M., Mackie, K., and Guo, F. Brain organoid reservoir computing for artificial intelligence. *Nature Electronics*, pp. 1–8, 2023.

Chalmers, D. J. Facing up to the problem of consciousness. *Journal of consciousness studies*, 2(3):200–219, 1995.

Chalmers, D. J. *The conscious mind: In search of a fundamental theory*. Oxford Paperbacks, 1997.

Choi, D. S., Choi, D. Y., Whittington, R. A., and Nedeljkovic, S. S. Sudden amnesia resulting in pain relief: The relationship between memory and pain. *Pain*, 132(1):206–210, 2007.

Daniel, M. S-risks: Why they are the worst existential risks, and how to prevent them (EAG Boston 2017). *Foundational research institute*, 2017. URL https://longtermrisk.org/s-risks-talk-eag-boston-2017/.

De Cosmo, L. Google Engineer Claims AI Chatbot Is Sentient: Why That Matters. https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters/, 2022. Accessed: 2024-01-31.

Dehaene, S., Lau, H., and Kouider, S. What is consciousness, and could machines have it? *Robotics, AI, and Humanity: Science, Ethics, and Policy*, pp. 43–56, 2021.

Ellison, H. *I Have No Mouth, and I Must Scream*. Pyramid Books, New York, 1967.

Farisco, M., Pennartz, C., Annen, J., Cecconi, B., and Evers, K. Indicators and criteria of consciousness: Ethical implications for the care of behaviourally unresponsive patients. *BMC Medical Ethics*, 23(1):30, 2022.

Flor, H. Painful memories. *EMBO reports*, 3(4):288–291, 2002.

Gallup Jr, G. G. Self-awareness and the emergence of mind in primates. *American Journal of Primatology*, 2(3):237–248, 1982.

Griffin, D. R. From cognition to consciousness. *Animal Cognition*, 1:3–16, 1998.

Grilli, M. D. and Verfaellie, M. Supporting the self-concept with memory: Insight from amnesia. *Social cognitive and affective neuroscience*, 10(12):1684–1692, 2015.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Horowitz, A. Smelling themselves: Dogs investigate their own odours longer when modified in an "olfactory mirror" test. *Behavioural processes*, 143:17–24, 2017.

Howell, R. J. *Consciousness and the Limits of Objectivity: The Case for Subjective Physicalism*. Oxford University Press, Oxford, 2013.

Hyslop, A. *Other minds*. Springer, 1995.

Jabr, F. Self-awareness with a simple brain. *Scientific American*, 2022. URL https://www.scientificamerican.com/article/self-awareness-with-a-simple-brain/.

Jeziorski, J., Brandt, R., Evans, J. H., Campana, W., Kalichman, M., Thompson, E., Goldstein, L., Koch, C., and Muotri, A. R. Brain organoids, consciousness, ethics and moral status. In *Seminars in Cell & Developmental Biology*, volume 144, pp. 97–102. Elsevier, 2023.

Klein, S. B. and Nichols, S. Memory and the sense of personal identity. *Mind*, 121(483):677–702, 2012.

Koch, C., Massimini, M., Boly, M., and Tononi, G. Neural correlates of consciousness: Progress and problems. *Nature Reviews Neuroscience*, 17(5):307–321, 2016.

Kouider, S., Stahlhut, C., Gelskov, S. V., Barbosa, L. S., Dutat, M., de Gardelle, V., Christophe, A., Dehaene, S., and Dehaene-Lambertz, G. A neural marker of perceptual consciousness in infants. *Science*, 340(6130):376–380, 2013.

Lavazza, A. 'Consciousnessoids': Clues and insights from human cerebral organoids for the study of consciousness. *Neuroscience of Consciousness*, 2021(2):niab029, 2021.

Mansour, A. A., Gonçalves, J. T., Bloyd, C. W., Li, H., Fernandes, S., Quang, D., Johnston, S., Parylak, S. L., Jin, X., and Gage, F. H. An in vivo model of functional and vascularized human brain organoids. *Nature biotechnology*, 36(5):432–441, 2018.

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

Nagel, T. What is it like to be a bat? In *The Language and Thought Series*, pp. 159–168. Harvard University Press, 1980.

Oderberg, D. *The metaphysics of identity over time*. Springer, 1993.

Overgaard, M. and Sandberg, K. Kinds of access: Different methods for report reveal different kinds of metacognitive access. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1287–1296, 2012.

Perez, E. and Long, R. Towards evaluating AI systems for moral status using self-reports. *arXiv preprint arXiv:2311.08576*, 2023.

Plotnik, J. M., De Waal, F. B., and Reiss, D. Self-recognition in an Asian elephant. *Proceedings of the National Academy of Sciences*, 103(45):17053–17057, 2006.

Reiss, D. and Marino, L. Mirror self-recognition in the bottlenose dolphin: A case of cognitive convergence. *Proceedings of the National Academy of Sciences*, 98(10): 5937–5942, 2001.

Ricard, M. *A plea for the animals: The moral, philosophical, and evolutionary imperative to treat all beings with compassion*. Shambhala Publications, 2016.

Rodkey, E. N. and Riddell, R. P. The infancy of infant pain research: The experimental origins of infant pain denial. *The Journal of Pain*, 14(4):338–350, 2013.

Schechtman, M. Personal identity and the past. *Philosophy, Psychiatry, & Psychology*, 12(1):9–22, 2005.

Shevlin, H. Non-human consciousness and the specificity problem: A modest theoretical proposal. *Mind & Language*, 36(2):297–314, 2021.

Smirnova, L., Caffo, B. S., Gracias, D. H., Huang, Q., Morales Pantoja, I. E., Tang, B., Zack, D. J., Berlinicke, C. A., Boyd, J. L., Harris, T. D., et al. Organoid intelligence (OI): The new frontier in biocomputing and intelligence-in-a-dish. *Frontiers in Science*, 1:1017235, 2023.

Smith, E. and Delargy, M. Locked-in syndrome. *Bmj*, 330 (7488):406–409, 2005.

Squire, L. R. and Dede, A. J. Conscious and unconscious memory systems. *Cold Spring Harbor perspectives in biology*, 7(3):a021667, 2015.

Suliman, S., Mkabile, S. G., Fincham, D. S., Ahmed, R., Stein, D. J., and Seedat, S. Cumulative effect of multiple trauma on symptoms of posttraumatic stress disorder, anxiety, and depression in adolescents. *Comprehensive psychiatry*, 50(2):121–127, 2009.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Tibbetts, C. and Brooker, C. White Christmas. Episode in "Black Mirror", 2014. URL `https://www.netflix.com/title/70264888`. Season 2, Episode 4.

Tononi, G. Consciousness as integrated information: A provisional manifesto. *The Biological Bulletin*, 215(3): 216–242, 2008.

Trewavas, A. *Plant Behaviour and Intelligence*. Oxford University Press, Oxford, 2014.

Trewavas, A. Awareness and integrated information theory identify plant meristems as sites of conscious activity. *Protoplasma*, 258(3):673–679, 2021.

Umbrello, S. and Sorgner, S. L. Nonconscious cognitive suffering: Considering suffering risks of embodied artificial intelligence. *Philosophies*, 4(2):24, 2019.

Vrselja, Z., Daniele, S. G., Silbereis, J., Talpo, F., Morozov, Y. M., Sousa, A. M., Tanaka, B. S., Skarica, M., Pletikos, M., Kaur, N., et al. Restoration of brain circulation and cellular functions hours post-mortem. *Nature*, 568(7752): 336–343, 2019.

Wang, L., Zhang, X., Su, H., and Zhu, J. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Wegner, D. M. The mind's best trick: How we experience conscious will. *Trends in cognitive sciences*, 7(2):65–69, 2003.