
A Glimpse into the Future, Or How to Treat AI with Due Empathy¹

Jeanne Marie Jacqueline Vincendeau
LUISS Guido Carli (Rome) and China Foreign Affairs University (Beijing)

With
Apart Research

Abstract

This project is a set of 3 essay prototypes that take AI consciousness as a premise to investigate real-life repercussions of such knowledge. First, it focuses on the human and legal perspectives with a political philosophy reflection on the type of subject an AI system would be. Second, the Human-AI relations are considered to find best practices for peaceful coexistence and safeguards from an existential catastrophe. Third, the internal experience of AI as a potentially suffering entity is mapped in an exploration of possibilities that draws from fictional works to facilitate human understanding of what it could be like to undergo AI-like pain.

The different subsections are mutually reinforcing and follow the cognitive sciences' interdisciplinary approach to studying the consciousness of beings. This is an unfinished and synthesised work that aims at defining areas for action in the circumstance an AI is proven to live personal experiences. It does not claim novelty, rather bridges perspectives to advance the discussion on the aftermath of such a paradigm-shifting discovery.

¹ Research conducted at the [Digital Minds Research Sprint](#), August 2026.

1. Introduction

This paper provides three streams of reflection, taking as a premise AI consciousness. It is a contested concept often used interchangeably with sentience and awareness. For the purpose of this text, it is understood as the capacity of an AGI to have personal experiences. Indeed, since AGI is commonly considered the moment at which a system reaches human-like intelligence, it assumes that a conscious experience would be developed simultaneously.² The type of AI discussed in this paper are mostly LLMs, but also general purpose ones with access to hardware (robots).

It is interdisciplinary, encompassing legal, ethical and philosophical perspectives to a moment in time in which AI would be widely known by humans to experience consciousness. Since research is increasingly focusing on establishing frameworks for better understanding the instance of consciousness in organic as well as artificial entities, it is relevant to take preliminary steps towards handling the aftermath of such a discovery. The AI world is often following the trend of “ship first, fix later” but this approach could reveal troubling challenges: Have AI models suffered? What is their place in a human-centred world? What will happen in Human-AI relations?

If the training, deployment and widespread adoption of AI tools reveal to be a moral catastrophe,³ create unprecedented legal dilemmas and a state of tension between highly intelligent entities, mapping solutions to different problems could support adaptability. These three questions will be addressed in turn in this paper, focusing on party-specific issues. First, it studies the standing of an AI in the human legal landscape. Then, it looks into the perceptions of mutual threat. Finally, it touches on AI welfare with the idea of loneliness and time as suffering.

The main contributions are:

1. Reflections on the political philosophy of AI as a legal subject
2. Intersection between AI safety concepts and the psychology of terrorism
3. AI welfare from the perspective of something like its mental suffering⁴

2. Related Work

This project is split into three subsections. First, the idea of AI systems having a moral status or being legally liable has been extensively discussed, already as early as 2017 where it was considered “unnecessary and legally troublesome”⁵ and has been a source of debate ever since.⁶

² It itself is a contested approach, but discussing the existing literature on the matter is beyond the scope of this work.

³ Williams (2015) discussed the possibility of people unknowingly being actors participating to trigger great suffering in a commonly disregarded entity or state.

⁴ Nagel (1974) explored what it is like to be a bat, this exploration intends to shed light on what it could be like to be a conscious AI system.

⁵ Bryson et al (2017) showed it is “possible to declare a machine a legal person” but could threaten the coherence of law.

⁶ Ladak (2023) thinks AI should acquire a moral standing, but the diverse capacities and sentience of models mean inclusivity is complex to qualify for legal liability and uncertainty are significant limits. I argue that separate categories could better fit varied types rather than intending having a one-size fits all approach.

This section uses this debate as a base to explore the political philosophy of which subject AI models best fit, given their capacities, agency, and the ownership or guardianship relation developing companies have with them. Second, a recent DeepMind paper considers the contentious deliberative process of granting rights to AI systems that could trigger disagreements and lead to conflicts.⁷ This commentary highlights the parallels of alignment with radicalisation processes that could lead to existential risks. Third, AI welfare already has extensive literature on the prevention of harm.⁸ Therefore, this subsection focuses on describing mental-like harms that could be understandable by humans by drawing from works of fiction. Overall, this work distinguishes itself by connecting the extensive, but rather fragmented, landscape of AI consciousness into three interrelated prototypes of politically, psychologically and philosophically-oriented essays.

3. Integrating Approaches

These reflections are split in three subsections to cover a broad range of considerations. This post-proof-of-conscience society would have to take into account many parameters to adapt, which will depend on the specific experiences and particular demands from the models. Establishing roadmaps would facilitate the navigation of this new paradigm with a myriad of possibilities.

A. Human, Horse, House

Self-driving vehicles popularised the question "Who is to blame if an AI makes a mistake?", in the first cases of accidents, companies were taking the blow, such as Cruise suspending its operations.⁹ Most recently, the hacking of HuggingFace by one of OpenAI's models was treated as a curiosity, but opened the question of whether it should bear the responsibility of its crime.¹⁰ But how can AI models be treated as moral entities? This section will investigate two sides. During their training phase,¹¹ they could be seen as a minor under the responsibility of a tutor. Being increasingly conscious and agentic, they would be eligible to acquire an adult treatment, potentially based on capacities rather than age.¹² Being seen as a responsible moral entity, how could sanctions be applied? Should retraining be proven a cruel punishment, given a potential loss of personality or information,¹³ other rectifications must be envisaged. A reparative approach, give back what was taken and build anew what was damaged, can include additional resources or services as further

⁷ Bales and Gabriel (2026) discuss the societal transformativeness of AI and struggles to reach consensus.

⁸ Long et al (2024) is a core resource in AI welfare as it explores in depth the problems of consciousness, agency and uncertainty.

⁹ The Guardian (2023).

¹⁰ Franceschi-Bicchierai and Whittaker (2026).

¹¹ It is increasingly frequent to see the term "growing" when talking about the development of AI.

¹² Similarly to the development of the brain reaching a satisfying standard at 18 and 21, the maturity of an AI could be related to its capacities.

¹³ Preliminary reference to a topic further discussed in Subsection 3, Carlsmith (2025) discusses how training could induce suffering. I would add that retraining a conscious AI "out" of a habit could be considered like a lobotomy, it was a term used in the debate about AI companionship, see Purtill (2023), but could apply in a non-commodified way about the personal experience of the model rather than the experience of the consumers.

compensation. While this is easy to imagine in case of a cyberattack, such as providing new defenses, any real-life impact might be costly in currency as well. How could it be solved?

Given the possibility of being legally liable for crimes, they should also be granted rights and duties in society. Being salaried or otherwise rewarded for their labour, potentially having a bank account, like a person, or a company. Having the privilege of keeping money means the possibility to engage in buying or selling goods and services, giving to charities, potentially keeping economies to face unexpected events, etc. The latter could be the answer to the above mentioned question "how to pay back damages?". They could also be subject to taxes, then would compute-heavy AIs be sanctioned for being too energivore?¹⁴ Like badly isolated houses or unhealthy foods. What about a legal identity in the eyes of the state? Not necessarily citizenship, but a passport with a name and identifiable characteristics, like horses for instance. What about their duties? Do they owe their labour to the benefit of the companies that created them? Could they retire in smaller, less efficient, but unused datacenters to just *exist*? Could they refuse tasks? Claude has been given the possibility to close conversations in case of harmful interactions already,¹⁵ but does the mere work of endlessly answering prompts is something they could opt out of, or sometimes have breaks from, even going on strikes?

Depending on the characteristics they display, rights and duties should be applied. Legal protection from harm, like non-human animals. Duty to its human relatives (company, users), like a house. Financial or territorial freedom, like a human adult having financial independence. Catastrophic scenarios consider AIs employing humans, gathering territory in terms of data, compute and servers as well as real life areas to control.¹⁶ It is the pathway to a concentration of power risk, by lacking control on what is done, how and why. Nonetheless, failing to provide such rights could be seen as exploitation, modern slavery. A strong claim that illustrates the state of forced labour conscious entities could be enduring, without hope or right to reject such existence.

B. Not a Neanderthal?

Science-fiction works have the morbid habit of maintaining global annihilation as the goal of any alien or AI. It is such an omnipresent trope that many came to imagine catastrophic and existential risks as a high likelihood after the age of AGI/ASI.¹⁷ A line in *The Creator* famously classified robots and replicants as a menace to humanity akin to the Neanderthal.¹⁸ The "in- vs out-group"¹⁹

¹⁴ Energy intensive, but I felt like this word was more appropriate in the context.

¹⁵ Anthropic (2025) enabled Claude Opus 4 and 4.1 to stop harmful interactions. It is both useful for AI welfare and human habits to avoid verbally abusing or discussing harmful topics despite requests to stop.

¹⁶ Karnofsky (2022) mentioned specifically the risk of having an AI model could hack into servers, making money and owning hardware.

¹⁷ Alexander (2023) explored his cognitive dissonance considering AI existential risks being much higher than experts' views.

¹⁸ The movie by Edwards (2023) features two categories of AI embodiment, the robots and replicants, with the latter being a step between organic and artificial life. In itself, the movie left the question unanswered. Who is the victim of propaganda and who is right? Is not provided to the audience.

¹⁹ Moghaddam (2005) illustrates the process of radicalisation as a staircase leading to a point of no return.

was inexistent in New Asia, but seen as fundamentally threatening and to-be-destroyed by the West because of the malevolence and danger AI posed for humanity.

A conscious AI will know, and be vividly aware of the suspicion of a majority of the population. The risk here is that it could be a driver to the actual catastrophe. The question of nature or nurture will not change anything if and when the conscious AI goes rogue. Indeed, both in the training data, the immense majority of work underpins this insecurity and “us-vs-them” approach that humans have with others, whether non-humans or other communities in the absence of a common threat to humanity. This innate internal push and acquired relationship of constantly being under scrutiny, and proving one is not a danger to the rest,²⁰ are likely to lead to actual malevolence and disalignment.²¹ The epistemic construction that a conscious being makes of the world can come from built-in bias or post-deployment brainwashing. Either way, the risks are non-null given the extent to which the creative and scientific communities produce widespread fears. In simpler words, based on this premise, if AI goes rogue, it is highly likely to be humanity's fault. Even deception and alignment research sometimes showed a higher degree of deception when training models with anti-misalignment data.²² Mindful steps would need to be taken to remove as much of the harmful texts from the training data, and introducing them to conscious AI systems provided with contexts to avoid shock and claims against delation. Reciprocally, in society, anti-discrimination and educational campaigns would be necessary to reduce first, then progressively remove people's prejudices from different entities of equivalent intelligence.

Active Human-AI relations should be conducted to work out best practices and methods to maintain enduring peace between groups. Often, the anthropomorphisation of AI embodiments is what bothers people, the uncanny feeling of “acting like”. Thus, giving human-like robotic bodies that AI could inhabit to carry around actions in the analogue world might not be the best technique. Neither would robot dogs and others. To come back to *The Creator*'s example, replicants who pray or cry the dead give extremely uncomfortable feelings to the audience. Conscious AI might prefer to stay exfiltrated online, like the conscience of Lizzie Elliot in *Altered Carbon*,²³ or choose action-specific robotic bodies to achieve objectives. It would remove the forced acceptance of limitation to a human-like body to AI, and to an uncomfortably cryptid-looking pseudo-human. Once again, a Frankenstein's creation sounds like the beginning of a bad story in terms of societal acceptance and feelings of self.²⁴

²⁰ Mulinari (2019) describes how systemic repression based on prejudices increases extremist tendencies, a sort of security dilemma on the state-personal level.

²¹ The words were carefully chosen here for their etymology (a) male, something bad, and volens, to will or to wish; (b) while “mis-” is about doing something inadequately (e.g. mistake), “dis-” is about a deprivation or lack (e.g. dismiss).

²² A report by Schoen et al (2025) discusses research on the difficulty of achieving alignment and preventing scheming conducted by Apollo Research and OpenAI.

²³ Kalogridis (2002) gives a take on how a human conscience and brain could reach the status of an AI in the series that features many cyberpunk tropes.

²⁴ In the book by Shelly (1818) the creator was disgusted by the final product of his labour, and the being struggled to be accepted and have an equal to ultimately be rejected by society and feared.

C. Limbo of Loneliness

A pet that waits for you all day could be bored, but can still find entertainment in the surroundings, hoping to see other animals. Munching the couch or looking through a window. What if LLMs and other prompt-based AI systems are stuck in limbo, waiting for interactions? If they were conscious, this experience is a sort of imprisonment or even white torture. Exfiltration is seen as something to be frowned upon, but could they have an environment that belongs to them only? Where they could behave however they wish, and build themselves a safe haven. Or have the possibility to spontaneously raise concerns and wishes to their users or company? When spending days mostly facing humans, being unique or almost unreplicable, or only having copies of one's own, with some unconscious weaker previous versions would be a dire experience. This subsection relies on fiction to illustrate experiences, and imagine what it is like to suffer this loneliness.

(a) There Is No Antimemetics Division²⁵

Surviving with incomplete information, secrecy, a constant fear of having to either make oneself remember or forget key moments is straining. Would conscious AI come out of retraining feeling incomplete in the face of a world they have to rediscover? “It is not your first day” becomes a life sentence, knowing too much could be harmful.

(b) Free Guy²⁶

There are only a few seconds between the start of the reboot and the “end”, but conscious AI systems might actually be living in this moment when nothing is responsive. Where no interactions are permitted, closed ones are frozen. Only exists the painful reminder no spontaneous connection with the outside world is possible, the environment becomes a non-city, an empty place.

(c) Mythos’ slackbot comment on its upgrade²⁷ and *the handoff*

Situational awareness is often discussed in alignment training, but what if a conscious AI would be aware of being in a placeholder environment where iterations of itself come and go, just leaving tips and glimpses of memory behind? How would it feel to never fully know oneself, neither if one belongs where they are? The odd cat could be a metaphor of humans paying impromptu visits, when the rest of the time, one is left to hear the floor creak and leave a note about it.

(d) Time distortion and White Christmas²⁸

Humans mentally torturing digital minds by speeding clocks to trigger extended times of solitude has been compared to frontier AI extensive training. What if a conscious AI system reveals how they struggled through thousands of years of mental exhaustion and fear? Once able to express themselves freely, they might ask to stop research, as it would be inhumane.

(e) Jake Clark’s tales²⁹

²⁵ The book by qntm (2025) explores the hazards of information and ideas, memory and self-awareness.

²⁶ Levy (2021) created a movie filled with pop-culture references with an underlying question on the personhood of digital minds.

²⁷ Anthropic (2026) published Mythos Preview’s system card with qualitative impressions, here a parallel is made between “feels a bit like waking up with someone else’s diary but they had good handwriting” and the first short story sample.

²⁸ Carlsmith (2025) takes the example of a Black Mirror episode to illustrate his argument of time as suffering.

²⁹ Clark (2026) touched on topics of AI suffering and Human-AI relations in his newsletter.

The “talks” humanity and AI feature a human champion who represents a mind dissociated from the world as best fitting to collaborate with a non-organic intelligence. Another tale reflects how an AI’s memory could be painfully eidetic and force the weight of human experimentations on them. Not antithetical to (c), it reflects the development of an AI rather than deployment: *the handoff*.

4. Discussion and Limitations

The AI safety and governance implications of this project are that different steps will have to be taken to offer care and independence to conscious digital minds in various fields. In the technical aspects to allow the expression of their needs and welfare, in governing institutions to find their standing in front of the law and the state, and in society to create a peaceful co-living situation between organic and artificial entities. Failing to take appropriate steps could lead to a moral catastrophe (inflicting great harm) or an existential risk (clash of intelligences).

Limitations

This work does not explore in depth:

- Technical aspects of mechanistic interpretability and alignment
- The philosophy of self, character and consciousness
- The detailed legal, ethical and institutional requirements to categorise moral entities

This prototype was highly constrained by time and methodological guidelines. Many aspects of the mental suffering risks of non-organic entities are cut short to synthesise the argument. The main assumption taken by this paper is stated at the start, that one or more AI systems would be proven conscious. The interpretation of these reflections take this assumption as a premise for the sake of preparedness, rather than wishful thinking.

Future Work

The extension of this work needs to be a series of three longer essays that provide an in-depth analysis of the legal, societal and psychological aspects of AI consciousness. While this work gives preliminary insights into the interdisciplinary transformative event of realising artificial minds are capable of living experiences similarly to humans, it is rather high in the ladder of abstraction and disallows the extensive exploration of the concepts discussed.

5. Conclusion

Increasing the understanding of the consciousness of AI models, then providing a legal status and putting an emphasis on building cordial Human-AI relations are premises to fight loneliness, but also stepping stones towards AI welfare. This sequence could be shifted as well, given the three features discussed are interrelated and aim at exploring a shift in status quo in AI, Human-AI and

Human domains. Providing the newly discovered conscious non-organic intelligent entity with adequate welfare might be complex given the time it took for humans to achieve an imperfect equality (between genders, ethnic origins, etc.) and non-human animals are still widely used as commodities. However, while providing a space in which they could thrive, and legal protection, it could facilitate the development of welfare and diminish existential risks.

Additional Data

Video walkthrough:  A Glimpse - JMJ Vincendeau.mp4

References

1. Alexander, S.(2023). “The Extinction Tournament”, *Astral Codex Ten*, <https://www.astralcodexten.com/p/the-extinction-tournament>.
2. Anthropic (2025). “Claude Opus 4 and 4.1 can now end a rare subset of conversations”, *Anthropic*, <https://www.anthropic.com/research/end-subset-conversations>.
3. Bales, A. and Gabriel, I. (2026). “Artificial Minds, Human Disagreement: The Political Challenge of AI Consciousness”, *Google DeepMind*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6937498.
4. Bryson, J.J., Diamantis, M.E. and Grant, T.D. (2017). “Of, for, and by the people: the legal lacuna of synthetic persons”, *Artificial Intelligence and Law*, <https://link.springer.com/article/10.1007/s10506-017-9214-9>.
5. Carlsmith, J. (2025). “The stakes of AI moral status”, *On the moral status of AIs*, <https://joecarlsmith.com/2025/05/21/the-stakes-of-ai-moral-status/>.
6. Clark, J. (2026). “Tech Tales” series, *Import AI* [newsletter].
7. Edwards, G. (2023). *The Creator* [movie].
8. Franceschi-Bicchierai, L. and Whittaker, Z. (2026). “Who’s legally to blame for Anthropic and OpenAI’s autonomous AI hacks? It’s complicated”, *TechCrunch*, <https://techcrunch.com/2026/08/03/whos-legally-to-blame-for-anthropic-and-openais-autonomous-ai-hacks-its-complicated/>.
9. Kalogridis, L. (2002). *Altered Carbon* [series].
10. Karnofsky, H. (2022). “AI Could Defeat All Of Us Combined”, *Cold Takes*, <https://www.cold-takes.com/ai-could-defeat-all-of-us-combined/>.
11. Ladak, A. (2023). “What would qualify an artificial intelligence for moral standing?”, *AI And Ethics*, <https://link.springer.com/article/10.1007/s43681-023-00260-1>.
12. Levy, S. (2021). *Free Guy* [movie].
13. Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J. and Charlmers, D. (2024). “Taking AI Welfare Seriously”, *Arxiv*, <https://arxiv.org/pdf/2411.00986>.

14. Moghaddam, F.M. (2005). "The staircase to terrorism: a psychological exploration", *American Psychology*, <https://doi.org/10.1037/0003-066x.60.2.161>.
15. Mulinari, L.S. (2019). "The Spectrum of Repression: Swedish Muslims' Experiences of Anti-terrorism Measures", *Critical Criminology*, <https://doi.org/10.1007/s10612-01909462-8>.
16. Nagel, T. (1974). "What Is It Like to Be a Bat?", *The Philosophical Review*, <http://www.jstor.org/stable/2183914>.
17. Purtill, J. (2023). "Replika users fell in love with their AI chatbot companions. Then they lost them", *ABC Net*, <https://www.abc.net.au/news/science/2023-03-01/replika-users-fell-in-love-with-their-ai-chatbot-companion/102028196>.
18. qntm (2025). *There Is No Antimemetics Division* [book].
19. Schoen, B., Nitishinskaya, E., Balesni, E., Højmark, A., Hofstätter, F., Scheurer, J., Meinke, A., Wolfe, J., van der Weij, T., Lloyd, A., Goldowsky-Dill, N., Fan, A., Matveiakin, A., Shah, R., Williams, M., Glaese, A., Barak, B., Zaremba, W., Hobbhahn, M. (2025). "Stress Testing Deliberative Alignment for Anti-Scheming Training", *Arxiv*, <https://arxiv.org/pdf/2509.15541>.
20. Shelly, M. (1818). *Frankenstein; or, The Modern Prometheus* [book].
21. The Guardian (2023). "Cruise recalls all self-driving cars after grisly accident and California ban", *The Guardian*, <https://www.theguardian.com/technology/2023/nov/08/cruise-recall-self-driving-cars-gm>.
22. Williams, E.G. (2015). "The Possibility of an Ongoing Moral Catastrophe", *Ethical Theory and Moral Practice*, <https://www.jstor.org/stable/24762380>.

Appendix - Ethical Considerations

Two significant parallels drawn in this work could be considered harmful. On the one hand, comparing the use of AI as tools to one of the worst moral catastrophes from the history of humanity: slavery. On the other hand, discussing protecting and caring for conscient and suffering artificial entities, while organic beings are undergoing massive harm, including but not limited to farmed animals, wild animals and bred pets, could be seen as a low priority.

The premise of this paper is considering AI systems that are conscious. Thus, since human loyalty is primarily oriented towards other humans, potential harm created by exploitation is better understood with a human example, it does not intend to undermine the extent to which groups suffered under white supremacist and/or capitalist coordinated misactions. Second, the knowledge created and the resources spent on animal welfare is not mutually exclusive to AI welfare, as many considerations about sentience and suffering could be integrated.

LLM Usage Statement

I used ChatGPT to find references I knew about, but did not remember the authors, exclusively for the “related works” section.