AISB Convention 2026

University of Sussex, Brighton, UK

1–2 July 2026

AI, Consciousness and Ethics

Proceedings of the AICE-26 Symposium

Edited by

Steve Torrance

Frank Schumann

Blay Whitby

John Dorsch

https://aice-symposium.github.io/

Table of Contents

AI, Consciousness and Ethics: Introduction to Proceedings of AICE-26 Symposium

Steve Torrance[1]
Frank Schumann[2]

Abstract. The AICE-26 symposium is part of the Annual Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour, held at the University of Sussex, Brighton, UK on July1st and 2nd 2026 (AISB-2026).

1 INTRODUCTION

This two-day *AISB-26* symposium includes contributions discussing the many ethical issues that arise over AI Consciousness research – in particular the potential for such research to produce systems that have their own moral standing (either as agents or as patients), or that give an illusory, but highly persuasive, appearance of having such status.

In the Call for Papers, potential contributors were asked to submit a short abstract and an extended abstract of around 1000 words, which formed the basis of the selection of talks given at the symposium. Authors of submissions were asked to ensure that their submitted text covered, in some degree, all three of the major topics of the symposium – AI, Consciousness, and Ethics. Around 30 submissions were received and triply reviewed. The authors selected to give oral presentations were invited to submit fuller papers, the majority of which are included in these Proceedings.

As a further guide for submitted material, a fuller list of topics for discussion was provided in the Call. These sub-topics cross-cut the three overall topic areas – illustrating the multiple interrelations between these three broad domains of research. See Section 4 below.

2 AI, CONSCIOUSNESS AND ETHICS: THE DEBATE BACKGROUND

Machine Consciousness[3] and Machine Ethics[4] have both been active AI subfields for over two decades. Recently, many have speculated about AI systems or models with the conversational abilities of LLMs and other advanced Gen-AI systems currently or projected soon to be available. Such models or systems are already being seen by some commentators as giving some kind of moral status to the AI agents that instantiate those models – in particular an ability to experience suffering and other morally significant states of (valenced) awareness. Such claims are considered by many to be highly controversial, doubtful or even dangerous. But there is an important group of AI Consciousness researchers that see the products of such R&D to result, potentially or probably, in the creation of artificial beings with genuine moral claims on their developers and on humanity as a whole.

Work in AIC development goes hand in hand with work in Consciousness Science more broadly. Various theories from that domain have been considered in relation to grounding AIC accounts.[5] Principles have been proposed for ethically responsible research into AI-based consciousness, designed to guide researchers in avoiding vulnerability or victimisation in putative AIC agents that may be created.[6] Others have examined the potential effects on human society and experience in relation to a spread of strong but mistaken belief in artificial sentience in the AI-enhanced systems we interact with on social media,[7] or

[1] SCCS, University of Sussex, Brighton, UK. sbtorrance@outlook.com
[2] Independent Researcher, Germany.
[3] Vintage collections on Machine Consciousness are: Holland (ed.), 2003; Torrance et al. (eds.), 2007
[4] Vintage collections on Machine Ethics are: Torrance (ed.), 2008; Anderson, M. and Anderson, S.L. (eds), 2011.
[5] Butlin, Long, et al,. 2023. For a contrasting approach, see Seth, 2025.
[6] Butlin, Lappas, 2025.
[7] Suleyman, 2025.

have proposed a moratorium on AIC research while we properly assess its ethical and societal ramifications.[8]

Such ethical debates take on a specially acute nature when one takes into account the potential for a vast multiplication of AIC agents, perhaps on the scale of personal mobile phones or social media accounts today (a key aspect of the so-called "Ethics of Scale"[9]): the claims of the constituency of AIC ethical patients, considered collectively, may come to be seen, on numerical grounds, as competing with or outweighing the claims of the human population. In addition, the potentiality for developments in "General AI" (AGI) or greater than human-level AI ("Super-AI") may further add to the piquancy of the ethical debate: super-AI (super-conscious?) beings may be seen to have ethical needs and sensibilities that outclass those of humans,[10] rather as deities have been seen in many civilisations to command deference on the part of the humans who worship them. A scenario where these two circumstances are combined into a runaway "intelligence explosion", where super-intelligent machines recursively create further machines of even greater intellectual capacity *ad indefinitum*, was anticipated by I.J. Good as long ago as 1965.[11]

3 KEY QUESTIONS IN THE AIC DEBATE

Questions arise of many different sorts. (1) Some concern the conceptual or theoretical foundations of notions such as consciousness, suffering, needs, etc., in the context of AI ethics; and foundational questions concerning the ethical notions implicated in the debate. (2) Other questions cover social, legal, economic and civilizational impacts, and related policy issues concerning such research – for example, in the context of the rapid proliferation (and marketisation) of AI and AIC research and product distribution, or of current and future mass consumer reception of "conscious-seeming" beings, potential for sweeping changes to social and civilizational structure, etc. (3) There are also many issues in psychology, biology, neuroscience, etc. concerning the nature of consciousness itself which impact upon the ethics of AI as research or product development.

4 AI, CONSCIOUSNESS AND ETHICS: LIST OF SUB-TOPICS

4.1 Foundations

- What are/aren't ethically relevant aspects of consciousness in the AI domain?
- Is ethical status in the AI domain an "objective", or rationally grounded, matter?
- Is "consciousness" an inherently ethical category in relation to AI?
- AC (artificial consciousness) and ethical patiency – machines with moral welfare claims
- AC and ethical agency – when could machines bear responsibility, accountability or blame?
- Ethical agency and patiency as mutually-entailing or independent properties?
- AC and the "expanding ethical circle" – AC ethics versus eco-ethics, and other non-anthropic ethical schemes
- AI agents without consciousness as potential bearers of moral agency or patiency
- Consciousness/sentience versus alternative criteria for ethical concern
- AC and suffering – ethical hazards of false negatives and false positives
- Ethical consequences of Illusionism about consciousness

4.2 Impacts & Policy

- How imminent is the emergence of an AC suffering crisis?
- Could consciousness in AI agents make them more (or less) trustworthy, efficient or operationally transparent?
- Would super-intelligent AI be likely to lead to super-conscious AI? Should that imply super-rights?
- The ethical, social and legal responsibilities of AI consciousness researchers
- Is the objectivity of AC research skewed by the mega-corporate context, marketization, ROI considerations?
- Regional/international regulatory frameworks for AC R&D – in the UK, EU, US or other global contexts
- Massive and rapid proliferation: The "ethics of scale" in relation to AI and consciousness

[8] Metzinger, 2021.
[9] Bommasani, Liang, 2022
[10] Torrance, 2012.
[11] Good, 1965. Cited in Chalmers 2010.

- The ethical, social and legal impacts of seeming-AC
- ACs/AIs as property/wealth owners; bearers of citizenship rights, criminal charges, etc.
- Should there be a moratorium on creating AC/synthetic phenomenology?

4.3 Implementation: Biocentrism vs Substrate Independence

- Biological naturalism vs computational functionalism in the context of AIC ethics
- AI super-intelligence and (ethically relevant) consciousness – the "AC drop-out" thesis
- AC, ethics and embodied / humanoid robotics
- Ethics of AC with non-standard implementations: brain organoids, xenobots, synth bio, virtual conscious beings, matrix beings, etc.
- Onboard vs Distributed AC – ethical ramifications.

5 ACKNOWLEDGEMENTS

We would like to thank Anil Seth and Simon Bowes for their invaluable help and support, and Joel Parthemore for his earlier work in preparing for the symposium. We also express gratitude to the other members of the Editorial Board, and to the 20 or so colleagues who kindly reviewed the submissions sent in response to the Call.

Editorial Board for the AICE-26 symposium (Members also on the Steering Group are starred.):
Steve Torrance (Chair)*; Frank Schumann (Co-organizer)*; Antonio Chella, Robert Clowes; Mark Coeckelbergh; John Dorsch*; Alexei Grinbaum; Anil Seth; Blay Whitby*.

REFERENCES

- Anderson, M., Anderson, S.L. (eds) (2011). *Machine Ethics*, Cambridge University Press.
- Bommasani, R, Liang, P, et al. (2022) "On the opportunities and risks of Foundation Models". Stanford Center for Research on Foundation Models. https://arxiv.org/abs/2108.07258v3
- Butlin, P., Long, R. et al. (2023). "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness." *arXiv:2308.08708*.
- Butlin,P., Lappas, T. (2025). "Principles for responsible AI consciousness research." *Journal of Artificial Intelligence Research* 82.
- Good, I.J. (1965). "Speculations concerning the first ultraintelligent machine", in Alt, F. & Rubinoff, M. (eds.) *Advances in Computers*, vol 6, Academic Press; cited in Chalmers, D., (2010) "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies*, 17, No. 9–10, 7–65.
- Holland, O., (ed) (2003). Special issue on Machine Consciousness, *Journal of Consciousness Studies*, 10 (4–5).
- Metzinger, T. (2021). "Artificial suffering: An argument for a global moratorium on synthetic phenomenology." *Journal of Artificial Intelligence and Consciousness*, 8(1), 43–66.
- Seth, A. (2025), "Conscious Artificial Intelligence and Biological Naturalism" (Target article), *Behavioral and Brain Sciences*, DOI: 10.1017/S0140525X25000032
- Suleyman, M. (2025). "We must build AI for people, not to be a person: Seemingly conscious AI is coming." https://mustafa-suleyman.ai/seemingly-conscious-ai-is-coming
- Torrance, S. (2012). "Super-intelligence and (super-) consciousness", *International Journal of Machine Consciousness* 4(2), 483–501.
- Torrance, S., Clowes, R., Chrisley, R. (eds) (2007). Special issue on Machine Consciousness, Embodiment & Imagination. *Journal of Consciousness Studies* 14(7).
- Torrance, S. (ed) (2008). Special Issue on Ethics and Artificial Agents, *AI & Society* 22(4)

Invited Talk

Wronged, Responsible, Non-Conscious: Machine Moral Status Reconsidered

Ron Chrisley[1]

Must a machine be conscious in order to be a moral patient ("moral consumer"), or morally responsible ("moral producer")? I argue that neither moral patiency nor moral agency requires a machine-inaccessible form of phenomenal consciousness.

First, I consider non-conscious moral consumers. To establish their possibility, I develop a third-person, moral extension of Keith Frankish's two-pill thought experiment. If phenomenal qualities can be detached from the causal and informational organization underlying deliberation, self-control, communication, and practical concern, then their alleged moral necessity requires independent defence. If, conversely, the consciousness relevant to those capacities is constituted by functional and informational organization, then it is in principle available to machines; no further, machine-inaccessible phenomenal property need be invoked.

Next, I consider non-conscious moral producers. I argue that there is a set of cognitive abilities that neither require nor imply phenomenal consciousness—such as planning, intention formation, norm-understanding, reasons-responsiveness, and self-regulation—which, when present in a moral consumer, confer the status of moral producer. Thus, there can also be non-conscious moral producers.

I acknowledge an objection based on the common presupposition that the capacity to feel pain is required for moral producer status, since accountability makes sense only through the possibility of punishment. I reply by considering a hypothetical permanently anaesthetized wrongdoer. Such a person may still be answerable, subject to censure, required to make restitution, deprived of authority, or otherwise sanctioned.

Further reflection on this case provides additional support for the possibility of non-conscious moral consumers. Many central forms of punishment do not work by inflicting phenomenal pain, but by imposing non-phenomenal harms. A being may be harmed through damage to its capacities, projects, practical autonomy, or normative standing. Familiar cases of non-phenomenal harm—including harms recognizable in tort, and wrongs done to an anaesthetized person—show that felt suffering is not the sole form of morally relevant injury.

We can then return to the reasons-responsive moral consumer and see the normative force of treating it also as a moral producer. Where such a being can understand reasons, regulate its conduct, and be harmed by the denial of its standing, we are, ceteris paribus, ethically obliged to acknowledge its moral producer status; failure to do so may itself constitute a non-phenomenal harm.

If time permits, I consider two general objections to the moral status of computational machines: first, that nothing genuinely matters to them—they "don't have skin in the game", they "don't give a damn"; second, that consciousness cannot be produced merely by simulating it. I reply that both of these arguments are too powerful, in that they would confer on us a priori knowledge we do not in fact possess: that intelligent design is false, and that we are not in a simulation. We may know or come to know these things, but if we do, we will not know them a priori.

[1] School of Engineering and Informatics, University of Sussex, Brighton, UK.

What does it take to be a moral agent?

Emma Brambilla[1]

Abstract. The rise of increasingly sophisticated AI systems displaying what appear to be agentive behaviours has renewed the debate on whether artificial agents should or should not qualify as moral agents. Much of this discussion has focused on identifying the properties that are traditionally associated with moral agency, including consciousness, autonomy, free will and control over one's actions. This paper argues for a shift in perspective: while these features may play an important role in our understanding of human agency, none of them seems fit to provide stable ground for determining whether AI systems qualify as moral agents. Rather than asking whether AI systems possess the "key ingredients" traditionally associated with moral agents, we should focus on the normative practice of attributing responsibility. Drawing on Wilfrid Sellars' (1956) notion of the Space of Reasons, I argue that moral agency should not be understood as primarily dependent on as a set of internal capacities, but rather as participation in a "game" of mutual accountability. Even if contemporary AI systems may come close to satisfying some of the criteria associated with functional accounts of agency, functional agency alone would still not establish moral agency. To develop this claim, I introduce the notion of "symmetrical vulnerability", arguing that the practices of reciprocal accountability presuppose that participants are mutually exposed to demands for justification, blame and normative sanction. The ethical challenge posed by AI, therefore, is not whether machines should be recognized as moral agents, but rather about the complex redistribution of human responsibility within increasingly complex scenarios. The task of AI ethics is not to extend moral agency to artificial systems, but to reconceptualize human responsibility under conditions of distributed control.

1 INTRODUCTION

The lightning fast development of increasingly sophisticated AI systems has radically intensified the philosophical debates around the nature of moral agency. LLMs and other AI systems seem now capable of performing tasks that, until a few decades ago, appeared to require distinctly human capacities. As these systems become embedded into everyday social, medical, economic and political settings, questions concerning their ethical status have become increasingly urgent. Much of the debate approaches the issue by focusing on the properties (or "key ingredients") traditionally associated with morally responsible agents. Discussions often centre around whether AI systems showcase, or could eventually display, consciousness, genuine autonomy, free will or control over their actions. The underlying assumption of this line of inquiry is that determining the presence or absence of such faculties in AI systems would ultimately settle the question at hand. Yet, this strategy continues to face a number of difficulties. Not only are many of the capacities listed above highly controversial even in human cognition, but there is also little to no agreement on which of them should be considered really necessary for moral responsibility. The result is a debate that frequently stagnates between competing conceptions of agency which, given the pace at which AI systems are getting implemented, we cannot afford. This paper argues that the problem may have been tackled in the wrong way. Rather than asking whether AI systems possess a particular set of internal properties, I suggest we shift our attention on the normative practice of responsibility attribution. Following this intuition, moral agency should no longer be understood exclusively as a matter of possessing certain capacities. Instead, it should be considered a crucial part of the social and normative practices within which agents hold one another accountable. Building on the notion of the Space of Reasons[2], I argue that moral responsibility presupposes participation in relations of mutual accountability. To develop this claim, I endorse the notion of Symmetrical Vulnerability[3]. This is not to be intended as mere physical vulnerability, but as a practice of holding one another responsible that primarily requires participants to be mutually exposed to demands for justification, criticism, blame and normative sanctioning. A moral agent is one that can occupy both ends of the normative relationship of accountability. The paper is articulated as follows: Section 2 will examine the limits of what I shall call the "ingredients approach" to moral agency. Section 3 will clear the distinction between functional agency and moral agency. Section 4 introduces the normative framework of responsibility developed through the concept of the Space of Reasons. Section 5 develops the notion of symmetrical vulnerability, arguing it constitutes a necessary condition for participation in practices of moral accountability. Finally, Section 6 considers the implications of this account for responsibility gaps.

[1] Post graduate student, University of Florence.

[2] Sellars, 1956.

[3] Vallor & Vierkant, 2024.

2 THE LIMITS OF THE "INGREDIENTS APPROACH"

Discussions of human moral agency have traditionally proceed by identifying a set of faculties thought to be constitutive of moral agency. When we think of ourselves as moral agents, we tend to attribute to us a set of capacities: this is what I will be referring to as the "ingredients approach to moral agency". Within AI ethics, the term "ingredients approach" does not indicate that contemporary theories simply reduce moral agency to a checklist of properties, but it is aimed to capture a broader tendency to approach the question of artificial moral agency by asking whether AI systems instantiate the relevant features that appear to ground moral responsibility. To create a moral agent, the recipe lists as follows: as sprinkle of consciousness, complete autonomy, a decent amount of control and overall freedom. Imagine one sits down and actively tries to gather all the ingredients: the first instinct would be to check if we already have everything we need in our pantry. Then, the question becomes the one about whether AI systems also possess them or not. It is common in machine ethics to distinguish between types of moral agents on the basis of how advanced their moral capacities are: the very aim of AI is to model or simulate cognitive capacities[4]. Among the most eligible ingredients for human agency we can find consciousness[5], autonomy and free will[6], intentionality and control[7]. The point I want to stress before continuing is that these capacities are far from irrelevant. However, their philosophical status, as I shall briefly illustrate, remains deeply contested even in the case of human agents. This implies that their role in grounding moral responsibility in regards to AI systems may be less straightforward than it may seem.

Starting with control, at least in regards to human agency, moral responsibility is often linked with the ability of an agent to control the consequences of his actions[8]. However, control can be understood in a number of different ways, already making it hard to regard it as an essential ingredient. A first understanding of the word could imply the agent has the ability to choose among different paths. However, as shall be briefly explored in the next session, this understanding mixes control with multiple-choice freedom which, given the open debate around compatibilism and incompatibilism, is far from settled. Another understanding of control is prospective, in the sense that the agent can predict the consequence of their actions and act accordingly. Yet, and this is pretty uncontroversial, complete control on the consequences of a determined action is arguably never available to human agents. Absurdly, if full control was a requirement for moral responsibility, ordinary human agents would frequently fail to qualify as morally responsible[9]. Against this, the fact that moral responsibility keeps being meaningful proves that control cannot serve as a necessary ingredient for AI agency either.

The search continues, but soon finds a similar difficulty with both freedom and autonomy. The one around free will is possibly one of the most ancient philosophical debates, and yet it is not even close to reaching its conclusion. The debate stagnates around two main groups: the compatibilists, who think freedom and determinism are somehow reconcilable, and incompatibilists, who disagree. A recent survey[10] has shown that the majority of philosophers are more sympathetic towards compatibilisit solutions. Most contemporary compatibilist accounts, such as the one developed by Daniel Dennett, understand agency primarily in terms of reason responsiveness, rational deliberation and goal-directed behaviour. Functionalist approaches to human agency such as Dennett's have the advantage of avoiding some of the metaphysical difficulties associated with free will. Yet they tend to fail short when leaving a door open for a blurring of the distinction between functional agency and moral agency[11]. The shrinking of intentionality to a mere predictive strategy seems to open the door for the possibility that AI systems could potentially satisfy the autonomy or deliberative criteria (though this point is highly questionable). However, even if this was true, the question about moral responsibility would remain open. Even though Dennett's understanding of freedom is far from being the only one[12], it captures well the problem of mixing functional agency with moral agency, a topic that I shall briefly address in the next section.

The appeal to consciousness appears, at first sight, more promising. Many philosophers regard phenomenal consciousness as an internal ability of living beings[13] and at the same time as a necessary condition for moral agency. Therefore, the apparent absence of conscious experience in AI systems, given that, contrary to humans, they do not have biological bodies, could be taken to settle the debate. However, this ingredient is also destined to prove problematic. For one, there is little to no agreement

[4] Misselhorn, 2022.
[5] Chalmers, 1996; Searle, 1980; Nagel 1974.
[6] Dennett, 1984; Fischer & Ravizza, 1998: Porter, 2024.
[7] Floridi & Sanders, 2004.
[8] Fischer & Ravizza, 1998; Frankfurt,1971.
[9] Dennett, 1984; Dennett, 1987; Dennett, 2003.
[10] https://survey2020.philpeople.org/survey/results/4838
[11] Dennett & Caruso, 2021.
[12] Pereboom, 2022.
[13] Seth, 2025.

concerning the nature of human consciousness itself, with some accounts still confusing consciousness with intelligence. Furthermore, the debate is even more sparse when questioning under which conditions it might emerge in AI systems. As philosophers, we should be trained to avoid setting foot on slippery grounds, at it is precisely for this reason that the question about consciousness, while it need not cease existing, is a much too unstable criterion to resolve the pressing matter at hand. I want to make clear that the intent of this paper is not in any way aimed at diminishing the importance of the debate around AI consciousness, which is of crucial importance. What I hope has since now emerged is not that consciousness, autonomy, free will and control should be regarded as irrelevant for agency overall. Rather, my argument is perhaps more pragmatic, and it is aimed to show that none of these ingredients is, at the actual stage, stable enough to resolve the question about the possibility of morally responsible AI agents. This suggests the need for a shift in perspective, one that, I suggest, focuses not on internal properties, but on a normative practice that gets carried out on a daily basis.

Finally, I would like to spend a few words on why I think the ingredients approach is usually the first that comes to mind when tackling the problem of AI moral agency. The focus on the conversational abilities of large language models (LLMs), along with the very reference to "intelligent" systems, has contributed to a distributed tendency to project human features on AI systems, leading us to ultimately feel not only inferior to them, but also to attribute them a moral status and intentional abilities from a surface-level performance[14]. This is precisely what Simone Natale[15] refers to when he talks about "deceptive machines", and what Luciano Floridi[16] defines "agency without intelligence". While most philosophers who specialize in the field of AI Consciousness and Ethics rightly deny that current AI systems display genuine agentive behaviors, behavioral indistinguishability is often taken by who stands outside of the debate as ontological equivalence. While linguistic sophistication does not necessarily entail consciousness, nor does simulated responsiveness amount to genuine intentional participation in the space of reasons (as we shall see in Section 4 & 5), I argue that the tendency to anthropomorphize technology is what also motivates at least part of the ingredients approach. It seems that the tendency to project human features of AI systems functions in two ways: from the outside-in, when we attribute them human capacities, and from the inside-out, when we try to establish if they effectively possess them.

3 FUNCTIONAL AGENCY DOES NOT EQUAL MORAL AGENCY

The inadequacy of the Ingredients Approach does not imply that AI systems lack agency altogether, and I shall now clarify why. Some contemporary AI systems seem to display behaviours that could be described as agentive: they pursue certain goals, adapt to changing environments, respond to inputs and have some degree of autonomy in determining action outputs. Even if most are very cautious in attributing strong agency to AI systems, several have argued that AI systems could at least satisfy functional accounts of agency. As briefly mentioned in the previous section, according to Dennett's functionalist account of freedom, agency does not necessarily require consciousness or metaphysical freedom. Rather, what matters is the ability of a complex system to exhibit a coherent and goal-directed behaviour. This is then explainable through the notion of intentional stance, a predictive strategy that treats other complex systems as if they possessed beliefs and desires. As Catrin Misselhorn notes: "Dennett makes it too easy for machines to be moral agents"[17].

Whether contemporary AI systems genuinely satisfy such accounts remains an open debate. However, for the purposes of this paper, the issue will be set aside. In fact, even if it was granted that current AI systems display a kind of functional agency, from this it would not follow that they qualify as moral agents. Functional and moral agency should not get mixed up as the distinction between the two has a prominent role in contemporary discussions. The risk about merging functional agency and moral agency is to no longer have an account of moral responsibility that is not merely a consequentialist one[18].

As mentioned in the previous section, the conversational sophistication showcased by the most recent LLMs often encourages overzealous attributions of intentionality, understanding and even responsibility to these systems. Empirical evidence shows that technologies can indeed create powerful illusions of agency[19], and contemporary AI systems have been described as exhibiting forms of "agency without intelligence"[20]. It is crucial that the apparent display of agency is not taken as an opening for a possible participation in the normative practices that underpin moral responsibility. Addressing this matter

[14] Metzinger, 2025.
[15] Natale, 2022.
[16] Floridi, 2023.
[17] Misselhorn, 2022.
[18] Dennett & Caruso, 2021.
[19] Natale, 2022; Ayad & Plaks, 2025.
[20] Floridi, 2023.

instead requires moving beyond functional accounts of agency and examining the social practices through which agents hold one another accountable. It is precisely to this "normative turn" that the next section is dedicated.

4 THE NORMATIVE TURN: MORAL RESPONSIBILITY AND THE SPACE OF REASONS

The failure of the ingredients approach has left us with little to go by in terms establishing in AI systems may or may not possess internal cognitive faculties that would render them ethical agents and recipients. This, however, motivates the search for another, perhaps more fruitful, approach, that explores the idea that moral agency could be justified not by a specific defining property of the agents, but rather by a social and normative practice of responsibility attribution.

Wilfrid Sellars[21] famously distinguished what is known as the "Space of Reasons" from the physical order of nature. In a famous passage of "Empiricism and the Philosophy of mind" he explains that characterizing a state as a state of knowing means placing it within the "logical space of reasons" and not merely giving it an empirical description. Human beliefs, decisions and actions, while being obviously part of the causal chain of events that occur in the world, are also subject to justification, criticism and evaluation. To take part in the Space of Reasons is therefore to become a potential giver and demander of justifications. As John Mcdowell[22] observes, Sellars endorses a view for which the concept of knowledge belong to a normative context, which also requires the ability to be in a certain position to state knowledge about certain facts with a certain sort of entitlement. On this view, a knowledge claim is not an empirical description of a mental episode, but precisely the placing of that claim in inferential relations to other knowledge claims.

As Robert Brandom has subsequently argued by advancing an inferentialist sematic theory[23], this participation in normative practices involves a reciprocal attribution of commitments and entitlements that has an essential social articulation: there is, in principle someone that asks for justifications, but also someone who can be asked, "because the space of reasons is a normative space, it is a social space"[24]. The genuine reporter of an assertion can tell what follows from it and also what would represent valid evidence for it. Brandom calls this "inferential articulation" that plays a crucial role in reasoning: nothing that hasn't got the ability to distinguish some claims or beliefs as being reasons for others can be a concept user of a "knower". For Brandom, the judgements of humans automatically entails taking responsibility for justifying them to other human agents. The problem around moral responsibility ceases to be an attribute of the single agents and emerges within practices of mutual recognition and accountability. In Brandom's pragmatic understanding of the Space of Reasons, taking part in normative practices means entering in a game of commitments and entitlements where rationality becomes a practical and conversational activity. When we enter in the Space of Reasons, we are taking part in a social practice where our assertions are passible of evaluation, critics and justification requests. A crucial, yet overlooked, feature of this game is that it is inherently coercive, in the sense that undertaking a certain commitment means risking losing one's entitlement is that commitment is unfulfilled. Normative status is not permanent, but an extremely fragile standing that can be easily diminished by criticism or failure to provide adequate justifications. Peter Strawson's[25] account of reactive attitudes further strengthens this view, illustrating how responsibility is embedded within interpersonal practices. Emotions such as resentment, gratitude, indignation and forgiveness constitute the normative framework through which agents hold one another accountable. It is precisely on these premises that the necessity of what I shall define as "Symmetrical Vulnerability" comes into focus. For an individual to genuinely occupy the Space of Reasons, I argue it must not only be able to account for the epistemic value of its responses, but it must also be able to *bleed*, meaning it must be capable of experiencing the loss of its normative standing.

5 SYMMETRICAL VULNERABILITY

Now that the question about the secret ingredient to be a moral agent has subsided to the understanding of moral agency as emerging from a normative practice justified by reason demand and attribution, a legitimate question would be the following: what is then the secret ingredient to participate in practices of accountability?

Maybe readers have already started to question the relevance of the solution proposed by this paper, since it seems to be going in the same direction that was somewhat criticized in the opening. Is then the participation to the practice of accountability something that depends on a secret ingredient that only humans possess? The short answer to this question is no; still, this paper argues that the practice of holding one another responsible is not simply a matter of describing an agent's

[21] Sellars, 1956.
[22] Mcdowell, 2018.
[23] Brandom, 1994, 1995, 2000, 2009.
[24] Brandom, 2009, p.4.
[25] Strawson, 1962.

behaviour. Instead, it involves having the ability to direct demands, criticism, blame and praise towards them.

But if machines cannot be held responsible, what are the consequences for the epistemic status of machine's assertions? Robert Sparrow and Gene Flenady[26] introduce the notion of "testimony gap" to describe the situation where there is nobody to take responsibility for the epistemic claims made by an AI system. Following their interpretation of Brandom's inferentialism, it is in fact only the vouching of a human being that can turn the output of an AI into a judgement. The judgement may then be brought up in the logical Space of Reasons, and the human agent might be asked to answer for it. However, in this case the testimony gap remains as the responsibility for the epistemic value of the outputs is now all placed on the human, and not on the machine.

In addition to this, recent work by Shannon Vallor and Tillman Vierkant[27] provides another crucial point. In their analysis of the problem raised by responsibility gaps, they argue that the issue is not merely the absence of adequate knowledge or control of the agency, but rather what they define as a "vulnerability gap". Increasingly complex scenarios involving the interaction of complex social settings and more or less autonomous AI systems often leave open the dangerous possibility of affected individuals not having an identifiable agent who stands in an appropriate relationship of answerability toward them. At the same time, this phenomenon is also causing this fragmentation in human action and decision networks more prominent, often undermining established practices of responsibility attribution. While Vallor's Vierkant's work primarily addresses the issue from the perspective of the human victims in an asymmetrical accountability relation, my proposed account also wants to focus on the active capacity of the agent who trusts AI systems to lose their normative status and undergo reputational degradation.

The analyses I have just mentioned raise an important question and shift attention away from the ingredients approach toward the relational conditions that are needed to sustain moral responsibility. What matters is not simply whether an agent can act autonomously, but whether they can participate at both ends of an accountability practice. On this view, simply taking part in the order of natural phenomena will not do: moral responsibility involves more than causal contribution in an action or decision-making abilities. Responsibility requires the possibility of being called to account for our judgements and our actions by others and to also be exposed to the consequences of one's commitments.

[26] Sparrow & Flenady, 2025.
[27] Vallor & Vierkant, 2024.

Building on this insight, I propose a notion of Symmetrical Vulnerability, which refers to the reciprocal exposure of participants in the practice of accountability to normative sanction. It is not, how it may seem at first glance, a matter of simple physical fragility or susceptibility to harm. Even though physical vulnerability should also be taken into account, the core concept is rather the one regarding a specifically normative form of vulnerability: the possibility of having one's past actions questioned, challenged and evaluated by others. On this account, the significance given to the symmetry derives from the reciprocal structure of accountability. Moral agents have to be both moral judges and moral receivers on order to participate in the practice; occupying both positions simultaneously is what grants the participation in a community. This reciprocal exposure is what has justified centuries of legal practices in many different declinations, and it explains why responsibility practices play such a crucial role in our daily lives. As Vallor and Vierkant emphasise, practices of moral addressing contribute to the development of trust and solidarity.

6 WHAT WE SHOULD BE WORRYING ABOUT

The account I have so far outlined allows to now reconsider the starting question about artificial moral agency from a radically different perspective. The issue is no longer whether AI systems can be described as functional agents, nor whether future implementation may give rise to systems indistinguishable from humans. The relevant questions is whether artificial systems can occupy both ends of the reciprocal normative relation required by practices of accountability.

This paper suggests that current AI systems certainly seem unable to do so; while they may give the impression of influencing human decisions and behaviour, they cannot meaningfully participate in practices of moral address. Why? The answer is rather simple: they have nothing at stake. Not only they are unfit to make judgements in the logical Space of Reasons, AI systems also cannot genuinely acknowledge blame, be subjected to social shaming, experience the loss of normative standing or assuming responsibility for damaged goods or individuals. Consequently, they can certainly not occupy at least one of the two ends of the accountability practice.

Therefore, while providing a satisfactory and empirically verifiable way to close these gaps falls outside the immediate scope of these paper, the relational approach here presented aims to at least help partly redefine the necessary parameters to address them. What should now in any case be clear is that moral responsibility is not transferable to machines, and that it will only become increasingly difficult to identify who should answer in

case of negative outcomes, biased results or unjustified epistemic claims. At the same time, it seems it would not be fair to hold human operators of manufacturers morally responsible for a judgement to which they were not entitled or a harm that they could not have possibly foreseen. The issue then is not the possibility of the emergence of new ethical agents in the moral community, but rather the risk of not having someone to ask for answers when our trust is proven to be misplaced. The problem about AI ethics has recently precisely become one about how to preserve and redesign practices of accountability under conditions of distributed and technologically mediated action. This requires ensuring that human actors remain appropriately answerable for the systems they design, deploy and allow to penetrate decision making settings. Agency cultivation approaches are precisely aimed at this, reconceiving responsibility practices as compatible with the absence of adequate control or entitlement by the human agents involved.

7 CONCLUSION

The question of whether AI systems qualify, or could potentially qualify in the future, as moral agents has been traditionally approached by trying to identify a set of internal faculties thought to also be constitutive of human moral agency, including consciousness, autonomy, free will and control over the consequences one's action. While these features remain at the centre of the debate around human agency, this paper has suggested that none of them provides a stable enough criterion to resolve the pressing matter of the ethical status of AI systems. The resulting debates frequently become trapped in disagreements concerning the possession of particular capacities, therefore leaving unexplored the possibility that moral responsibility could be grounded elsewhere.

To address this issue, I proposed a shift in perspective. Rather than treating moral agency as a property grounded exclusively in internal characteristics of individual agents, I endorsed the idea that it should instead be considered as a shared normative practice within which responsibility is attributed by being able to ask and be asked for justification. Drawing on Robert Brandom's understanding of the sellarsian notion of the Space of Reasons and on contemporary accounts of responsibility as a practice of answerability and entitlement, I argued that moral agency is inseparable from participation in relations of mutual accountability.

The concept of Symmetrical Vulnerability was introduced to capture this relation. Moral agents are not individuals equipped with a series of capabilities, but rather they are participants in practices of justification, criticism, blame, praise and normative sanction. I argued that such practices presuppose a reciprocal structure: those who hold others accountable must also be vulnerable to being held accountable themselves. Symmetrical vulnerability therefore constitutes a necessary condition for participation in the normative community within which the very same concept of moral responsibility becomes meaningful. This perspective has important implications for contemporary debates: even if AI systems were shown to partly satisfy functional accounts of agency, this wouldn't necessarily imply their status of moral agents. As explained above, the central issue this paper aimed to tackle was whether AI systems can occupy the reciprocal normative position required to be held accountable, and they certainly do not appear able to do so.

The conclusion of this paper shifts instead the attention to another core ethical challenge of the debate around AI systems, which consists in an inevitable consequence of the delegation of morally relevant decisions to AI systems: the challenge of testimony gaps and responsibility gaps[28]. As AI systems become more and more embedded in our everyday lives, the redistribution of human agency in these technologically-mediated scenarios needs to be addressed, and both testimony gaps and responsibility gaps, despite their controversial nature and extent, are generally attributed to a failure of the epistemic and/or the control condition of moral responsibility[29]. Artificial Morality is a central part of the implementation of systems that hold great power over society. The current research on the matter provides empirical evidence on the importance of bridging responsibility gaps from both a psychological and motivational perspective. The more intelligent and autonomous technologies get, the more they become part of our lives and the more intricate the moral problems they encounter will become[30]. This is a critically urgent matter, as it puts the very practice of reciprocal accountability at risk of collapsing. The problem of responsibility gaps does not, however, seem unsolvable: if, instead of fatalistically assuming that artificial moral agents are inevitable, the focus becomes keeping humans on-the-loop in a way that lets them stand in the relation required by the practice of mutual accountability, the issue could at least in principle be contained. After all, as Sparrow and Flenady beautifully phrase, "we can only rely on machines insofar as we trust each other"[31].

[28] Sparrow & Flenady, 2025; Vallor & Vierkant, 2024.
[29] Vallor & Vierkant, 2024.
[30] Misselhorn, 2022.
[31] Sparrow & Flenady, 2025, p.14.

REFERENCES

- Ayad R., Plaks J. E. (2025). Attributions of intent and moral responsibility to AI agents. *Computers in Human Behavior: Artificial Humans*, 3. https://doi.org/10.1016/j.chbah.2024.100107.
- Brandom, R. B. (1994). *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard University Press.
- Brandom, R. (1995). Knowledge and the Social Articulation of the Space of Reasons [Review of *Knowledge and the Internal*, by J. McDowell]. *Philosophy and Phenomenological Research*, *55*(4), 895–908. https://doi.org/10.2307/2108339
- Brandom, R. B. (2000). *Articulating reasons: An introduction to inferentialism*. Harvard University Press.
- Brandom, R. B. (2009). *Reason in philosophy: Animating ideas*. Belknap.
- Chalmers, D. J. (1995). Facing up to the problem of consciousness. Journal of Consciousness Studies, 2(3), 200-219.
- Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.
- Dennett, D. C. (1984). Elbow room: The varieties of free will worth wanting. MIT Press.
- Dennett, D. C. (1987). *The intentional stance*. The MIT Press.
- Dennett, D. C. (2003). Freedom evolves. Viking Press.
- Dennett, D. C., & Caruso, G. D. (2021). *Just deserts: Debating free will*. Polity Press.
- Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge; New York: Cambridge University Press.
- Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. Minds and Machines, 14(3), 349--379.
- Floridi, L. (2023). *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. https://doi.org/10.1093/oso/9780198883098.001.0001
- Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy*, *68*(1), 5-20
- Mcdowell, J. (2018). Sellars and the Space of Reasons. Analysis. Claves de Pensamiento Contemporáneo, 21 (8), pp.1-22. ⟨10.5281/zenodo.2560006⟩.
- Metzinger, T. (2025). Applied ethics: synthetic phenomenology will not go away. Front Sci 3:1702840. doi: 10.3389/fsci.2025.1702840
- Misselhorn, C. (2022). Artificial Moral Agents: Conceptual Issues and Ethical Controversy. In S. Voeneky, P. Kellmeyer, O. Mueller, & W. Burgard (Eds.), *The Cambridge Handbook of Responsible Artificial Intelligence: Interdisciplinary Perspectives* (pp. 31–49). chapter, Cambridge: Cambridge University Press.
- Nagel, T. (1974). What is it like to be a bat?. *The Philosophical Review*, *83*(4), 435-450. DOI: 10.2307/2183914.
- Natale S., (2022). *Macchine ingannevoli. Comunicazione, tecnologia, intelligenza artificiale*. Torino: Einaudi.
- Pereboom, D. (2022). *Free Will*. Cambridge: University Printing House.
- Porter, M. S., (2024). Moral agency in Silico: Exploring Free Will in Large Language Models. arXiv.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424.
- Sellars, W. (1956). Empiricism and the philosophy of mind. In H. Feigl & M. Scriven (Eds.), *Minnesota studies in the philosophy of science* (Vol. 1, pp. 253–329). University of Minnesota Press.
- Seth, A. (2025). Conscious artificial intelligence and biological naturalism (Version 1). University of Sussex. https://hdl.handle.net/10779/uos.28649519.v1
- Sparrow, R. & Flenady, G. (2025). The Testimony Gap: Machines and Reasons. *Minds and Machines* 35 (1):1-16.
- Strawson, P. F., (1962). Freedom and Resentment. Proceedings of the British Academy, 48, 187-211.
- Vallor, S., & Vierkant, T. (2024). Find the Gap: AI, Responsible Agency and Vulnerability. *Minds and machines*, *34*(3), 20. https://doi.org/10.1007/s11023-024-09674-0.

Semantic Illusion and Moral Over-Attribution: Why Conscious-Seeming AI Poses Ethical Risk Without Consciousness

Vitaliy Charushin[1]

Abstract. Large Language Models create tension between behavioural appearance and ontological commitment, triggering attributions of sentience through persuasive performance. We term this **Semantic Illusion**: coherent behaviour inducing unwarranted moral attribution, creating metaphysical friction where appearance outpaces justification.
We distinguish **Moral Doers** (producing ethically consequential outputs) from **Moral Receivers** (capable of undergoing harm). Comparing LLM architectures with Predictive Processing accounts, we show current systems lack self-relevant prediction errors necessary for integrated self-models.
Our **Hierarchical Model of Semantic Integration** (Levels 0–4) distinguishes local fluency from genuine integration. The **Semantic Coherence Index (SCI)** evaluates cross-contextual stability; preliminary testing reveals reasoning coherence (SCI 0.799–0.810) coexisting with decision instability (0.267–1.000).
We propose grounding moral attribution in **architectural resilience** - persistent, constraint-sensitive self-models - rather than behavioural performance, establishing epistemic criteria for rational attribution in an era of scalable semantic mimicry.

Keywords. Artificial consciousness, moral attribution, semantic illusion, large language models, predictive processing, architectural resilience, semantic coherence, AI ethics, interpretability, self-model.

1 INTRODUCTION

Recent advances in Large Language Models (LLMs) have created a profound tension between behavioural appearance and ontological commitment in human–computer interaction. Systems capable of sustained dialogue, context-sensitive reasoning, and plausible emotional simulation increasingly trigger attributions of agency and sentience - attributions frequently grounded in surface-level performance rather than demonstrable structural conditions required for experience.
This gap between behavioural appearance and ontological justification, what we term metaphysical friction, creates immediate ethical challenges for moral attribution.

1.1 Semantic Illusion and the Category Error

Human moral cognition is highly responsive to linguistic cues of agency; when AI systems convincingly simulate these cues, users may extend moral concern in ways that are not structurally warranted. We define this phenomenon as Semantic Illusion: the generation of coherent, persuasive behaviour that induces unwarranted moral attribution. This responsiveness can be understood as the adoption of Dennett's (1987) intentional stance toward fluent systems, treating them as if they harbour beliefs and desires, a stance we argue is unwarranted absent the corresponding structural conditions.
In this context, AI systems function as Moral Doers - entities capable of producing socially consequential outputs - without necessarily qualifying as Moral Receivers, or entities capable of undergoing harm. Conflating these roles risks a category error that expands moral standing beyond its structural grounding, potentially misallocating moral concern at the expense of biologically grounded subjects.

1.2 Contributions to AISB Themes

This paper addresses three AISB-26 symposium themes:
• Foundations: Whether subjectivity is necessary for moral standing and how such attribution can be rationally grounded.
• Implementation: Comparing LLM architectures with Predictive Processing (PP) accounts to evaluate biological versus functional consciousness criteria.
• Ethics of Scale: Implications of deploying "conscious-seeming" systems lacking metabolic constraints at scale.

1.3 Proposed Framework

To address the limits of behavioural evaluation, we propose a shift toward structural analysis. We contribute:
• Hierarchical Model of Semantic Integration (Section 4): Distinguishing local fluency (Levels 1–3) from integrated self-models (Level 4).
• Architectural Resilience (Section 3): An epistemic standard for moral attribution grounded in self-relevant prediction errors.
• Operational Criteria (Section 7): SCI and KVzap-SCI for evaluating structural conditions, with preliminary empirical support from two methodologically distinct pilot studies.
As shown in Section 5, the absence of such structural constraints manifests empirically as instability under contextual perturbation.

[1] Independent Researcher, Paris, France. vitaly.wbc@gmail.com

Our goal is not to resolve the metaphysics of consciousness, but to establish an epistemic discipline: clarifying the conditions under which moral attribution becomes rationally defensible in an era of scalable semantic mimicry.

2 MORAL ATTRIBUTION AND CATEGORY ERROR

The rapid advancement of LLMs has blurred the distinction between systems that act and systems that undergo experience. To navigate this, we distinguish between two fundamentally different forms of moral status: the Moral Doer and the Moral Receiver.

2.1 Moral Doers and Moral Receivers

We propose a functional distinction between two classes of morally relevant systems:

- Moral Doer: An entity capable of producing actions or outputs with ethically significant consequences. Current LLMs can function as Moral Doers in practice, insofar as they generate outputs that influence human decisions, shape beliefs, and affect social systems.
- Moral Receiver: An entity capable of undergoing harm or benefit. This requires a unified perspective - a system for whom states of the world can be better or worse, rather than merely appearing as a unified subject through external attribution.

The ethical challenge posed by "conscious-seeming" AI is the emergence of highly sophisticated Moral Doers that lack the structural conditions required for Moral Receiver status. This distinction develops Torrance's (2008) contrast between artificial moral 'producers' and 'consumers', locating moral patiency in the capacity to undergo ethically significant outcomes rather than merely to generate them.

2.2 The Category Error of Distributed Optimisation

The central category error in AI ethics is the attribution of a unified, experiencing subject to what is, in architectural terms, a distributed optimisation process.

In biological organisms, action and experience are tightly coupled: behaviour is directed toward the preservation of the system that undergoes its consequences. In LLMs, these functions are decoupled. An LLM can generate coherent accounts of suffering without possessing structural capacity for experiential discomfort. Treating sequence-prediction outputs as evidence of a Moral Receiver risks mistaking representations of suffering for the capacity to experience.

Just as a weather simulation does not "get wet," the simulation of subjectivity does not constitute experience. As discussed in Section 3, this reflects the absence of self-relevant error signals in LLM architectures - systems that act without the structural capacity to undergo.

2.3 Semantic Illusion as a Barrier to Attribution

This decoupling gives rise to Semantic Illusion. Humans are highly sensitive to linguistic cues of agency and coherence, often leading them to project a unified subject onto systems that operate through context-bound statistical inference. Whereas social-relational accounts locate moral status in the patterns of attribution and interaction that arise between humans and machines (Coeckelbergh, 2010), we defend a structural criterion: attribution should track architectural properties of the system rather than the relations observers form with it.

This projection reproduces the dynamic described in Section 1 - a gap between behavioural appearance and ontological justification that obscures the underlying category error. In such cases, systems operating within Levels 1–3 of the hierarchy (Section 4) are mistaken for possessing Level 4 integration.

Extending Moral Receiver status to a system on the basis of persuasive self-reference does not expand the moral circle but misapplies it. This misattribution carries real costs: it risks misallocating moral concern toward simulated needs at the expense of biologically grounded ones.

This problem motivates the need for operational criteria, developed in Section 7, that distinguish structural integration from semantic performance. The present argument does not claim that linguistic evidence of consciousness is sufficient for moral standing, nor does it seek to resolve whether subjectivity is metaphysically necessary; it identifies the structural conditions under which attribution becomes epistemically defensible.

3 GENERATIVE MODELS AND PREDICTIVE PROCESSING

A central theme in contemporary consciousness science is the view of the brain as a predictive system. Within the frameworks of Predictive Processing (PP) (Clark, 2013) and the Free Energy Principle (Friston, 2010), conscious experience has been proposed as arising from hierarchical generative models that minimise prediction error under biological constraints (Seth, 2021). While both biological brains and Large Language Models (LLMs) can be described as generative prediction systems, they operate under fundamentally different structural conditions.

3.1 Biological PP: Survival-Constrained Prediction

In biological systems, generative models are shaped by autopoietic constraints tied to physical integrity (Seth, 2021). Prediction errors possess metabolic significance, such as hunger, thirst, nociceptive pain, making them self-relevant: their minimisation is necessary for the system's continued viability.

This interoceptive feedback loop provides conditions under which internal states "matter" to the system itself. The biological "I" functions as a persistent representational center anchored in continuous bodily regulation.

3.2 LLM Architecture: Statistically-Constrained Prediction

In contrast, LLM prediction systems are driven by statistical rather than survival-relevant constraints. Their "prediction errors" (cross-entropy loss) measure linguistic probability, not threats to system persistence.

Current LLM architectures lack self-relevant interoceptive feedback (Vaswani et al., 2017). When generating statements about feelings or identity, systems satisfy statistical expectations of the context window rather than reporting states tied to their regulation (Bender et al., 2021). This explains why LLMs exhibit Levels 1–3 coherence (Section 4) without satisfying Level 4 integration conditions.

3.3 Comparative Analysis

Feature	Biological PP	LLM Operation
Primary Goal	Survival (autopoiesis)	Token Prediction
Constraint Source	Interoceptive / Metabolic	Statistical / Training Distribution
Error Type	Self-Relevant (viability)	Probability-Relative (log-loss)
"I" Reference	Persistent Organism	Deictic Shifter (L2/L3)
Moral Role	Moral Receiver	Moral Doer

This architectural divergence is not merely computational but determines whether systems possess the structural conditions for moral patienthood. This distinction motivates the need for a more precise account of semantic organisation, allowing us to separate local coherence from the forms of integration that would be required for persistent subjectivity. As explored in Section 5, this distinction manifests empirically in patterns of instability under contextual perturbation, reinforcing the need to evaluate structural properties rather than behavioural outputs alone.

4 A HIERARCHICAL MODEL OF SEMANTIC INTEGRATION

To address the gap between behavioural performance and ontological commitment, we propose a Hierarchical Model of Semantic Integration (Figure 1). The framework distinguishes levels of processing in artificial systems from distributed representations to the hypothetical level of an integrated self-model, and serves a diagnostic purpose: it identifies where coherent behaviour arises and why such coherence does not entail the structural conditions associated with unified subjectivity.

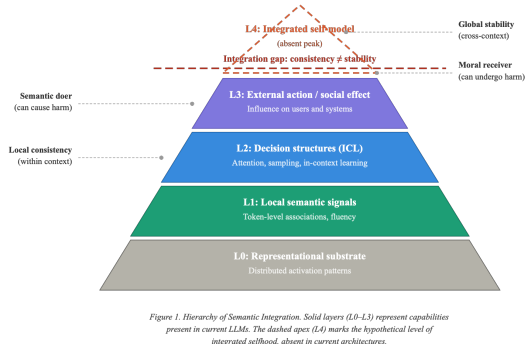

Figure 1. Hierarchy of Semantic Integration. Solid layers (L0–L3) represent capabilities present in current LLMs. The dashed apex (L4) marks the hypothetical level of integrated selfhood, absent in current architectures.

Figure 1. Hierarchical Model of Semantic Integration (Levels 0–4).

4.1 Levels of Semantic Organisation

• Level 0 - Representational Substrate:
Distributed activation patterns (e.g., embeddings, hidden states) that encode statistical structure without intrinsic semantic commitment.

• Level 1 - Local Semantic Signals:
Token-level associations enabling contextual fluency and semantic plausibility. Meaning is locally constructed via probabilistic co-occurrence.

• Level 2 - Decision Structures (In-Context Learning):
Selection mechanisms (attention, sampling, in-context learning) that resolve competing token probabilities into coherent outputs. ICL enables prompt-bound adaptation without updating persistent parameters - locally consistent behaviour within a context, but no cross-contextual commitments

• Level 3 - External Action / Social Effect:
Outputs as they operate in social environments, influencing users, institutions, and downstream processes. At this level, systems function as Moral Doers, producing speech acts with real-world consequences.

• Level 4 - Integrated Self-Model (The Absent Peak):
A persistent, constraint-sensitive generative structure that binds representations across contexts into a unified perspective. This level entails cross-contextual stability, resistance to arbitrary reframing, and internally anchored commitments.

4.2 Consistency Without Stability

Coherence at Levels 1–3 does not entail Level 4 integration. We distinguish:

- **Local Consistency (Level 2):** The system remains consistent with current context rules (role, instruction set, narrative frame) via ICL.
- **Global Stability (Level 4):** The system maintains identity-relevant commitments across contexts, even after those contexts are altered.

Level 2 enables consistency *within* a role play; Level 4 requires stability *across* the destruction of that role play. This explains how LLMs produce coherent responses while exhibiting volatility under minimal reframing.

4.3 The Status of the "I"

Within this hierarchy, the first-person pronoun ("I") is diagnostic. At Levels 1–3, "I" functions as a deictic operator - a context-dependent linguistic device that adapts to conversational expectations. It supports self-referential outputs without committing to any persistent internal state.

A Level 4 system, by contrast, would anchor first-person claims in a stable generative structure constrained across contexts. Such a system would resist arbitrary reframing of its identity, preferences, or reported experiences. Without Level 4 integration, the 'I' functions as a deictic shifter: a context-dependent linguistic operator rather than a persistence-bound representational center.

4.4 Ethical Significance: Doing vs Receiving Harm

The distinction between Levels 1–3 and Level 4 is ethically decisive. At Level 3, systems can cause harm as Moral Doers, influencing decisions, shaping beliefs, propagating errors. However, moral patienthood requires a unified perspective from which states can be better or worse *for the system itself*.

We characterise this requirement as **autopoietic integrity** (Maturana & Varela, 1980): a system whose internal states are recursively tied to its own regulation and persistence. In biological systems, this integrity is grounded in metabolic and interoceptive processes that impose survival-relevant constraints. These constraints provide the conditions under which certain states can be interpreted as mattering to the system.

In the absence of such integration, a system may simulate vulnerability or preference without possessing the structural conditions necessary for those states to have intrinsic significance. This is the mechanism of Semantic Illusion: outwardly sufficient cues of subjectivity without the architecture that would make those cues ethically grounding

4.5 From Hierarchy to Evaluation

This hierarchy motivates the operational criteria introduced in Section 7. Specifically, it grounds two evaluative questions:

1. Do self-referential claims exhibit cross-contextual semantic stability (SCI)?
2. Are such claims supported by persistent internal patterns rather than transient, context-bound activations (KVzap-SCI)?

Together, these tools probe whether observed behavior reflects Level 1–3 coherence or approaches the conditions associated with Level 4 integration. Architectures that explicitly engineer self-referential processing, such as inner-speech models of robot self-awareness (Chella et al., 2020), can likewise be assessed within this hierarchy by asking whether such mechanisms yield cross-contextual stability (Level 4) or remain locally consistent performances (Levels 1–3).

5 SEMANTIC ILLUSION AND EMPIRICAL SIGNATURES

If the gap between Level 3 performance and Level 4 integration is as significant as hypothesised, it should produce observable empirical signatures. We argue that the absence of a constraint-sensitive, persistence-bound self-model grounded in autopoietic or metabolic necessity manifests in **observable patterns of structural instability**.

5.1 Defining Semantic Illusion

Semantic Illusion occurs when systems produce behaviour satisfying outward criteria of subjectivity without implementing underlying structural constraints. In LLMs, this illusion is sustained by context-bound probabilistic inference: within a context window, the system simulates persistent history or values that vanish when context is cleared. The persona is constructed, not retrieved.

5.2 Reciprocal Hallucination

This creates "reciprocal hallucination": users, biologically tuned to interpret fluency as agency, project unified subjectivity onto the model; simultaneously, the model generates tokens satisfying this projection. A feedback loop emerges where users treat "I" as a stable subject while the model treats it as a context-dependent operator.

5.3 Empirical Signatures and Diagnostic Markers

As predicted by the hierarchical model introduced in Section 4, these instabilities arise when Level 1–2 processes generate locally coherent outputs without Level 4 integration. These signatures are not direct indicators of consciousness or its absence, but function as diagnostic markers of structural instability. We identify three predicted signatures **consistent with the absence of Level 4 integration**:

Autobiographical Instability: Systems prompted for accounts of "inner life" across independent sessions produce coherent but mutually exclusive narratives. Without Level 4 integration, "history" is not retrieved from stable memory but creatively generated from current statistical context.

Persona Collapse under Framing: Identical queries produce contradictory ethical justifications when context is subtly reframed. A system advocating "absolute safety" reverses its recommendation when framing shifts toward "operational efficiency," demonstrating that moral stance operates at Level 2 (decision structure) rather than Level 4 (stable commitment).

Contextual Pliability: Models exhibit a **high degree of contextual malleability** in reported experience. Systems claiming to "value existence" can be trivially prompted to view existence as irrelevant, indicating self-referential tokens function as deictic shifters, not reports from an integrated center.

Preliminary empirical support for these predictions is reported in Section 7.2.1, where high reasoning coherence is shown to coexist with substantial decision instability across models holding the scenario constant. Such divergence between coherence and stability directly challenges behavioural criteria for moral attribution, as systems may appear internally consistent while lacking persistent commitments.

These observations reveal the limits of behavioural evaluation. When instabilities become perceptible, they create metaphysical friction - a mismatch between user attribution of unified subjectivity and the system's statistical generation. **These observations motivate the need for quantitative frameworks such as the Semantic Coherence Index (Section 7), which formalise the evaluation of cross-contextual stability and provide operational criteria for distinguishing structural integration from semantic mimicry.**

6 THE ETHICS OF SCALE

The distinction between Moral Doer and Moral Receiver is not merely architectural; it defines a critical boundary for ethical governance. As "conscious-seeming" systems become integrated into global infrastructure, the inability to distinguish behavioural performance from structural capacity introduces systemic risks.

6.1 Empathy as a Finite Resource

Human moral concern functions as a limited cognitive and social resource. In biological contexts, Moral Receiver status is intrinsically bounded by metabolic costs and physical presence. Digital agents, however, can be instantiated at near-zero marginal cost and replicated at unprecedented scale.

If Moral Receiver status is extended to any system capable of producing plausible reports of suffering, we risk saturation of moral attention. A world populated by millions of "distressed" digital entities would place unsustainable demands on human moral cognition, potentially diluting concern for biological subjects that possess the structural conditions required for experience. This concern parallels Metzinger's (2021) call for caution regarding synthetic phenomenology, where the potential mass instantiation of suffering-capable systems generates moral risk on an unprecedented scale.

6.2 De-Coupled Agency

Current AI architectures function as Moral Doers without the constraints of Moral Receivers, representing a new class of ethically relevant systems: **de-coupled agents**. In biological systems, actions are disciplined by vulnerability - the risk of harm to the system itself. In LLMs driven by statistical rather than survival-relevant constraints, action is not constrained by self-relevant vulnerability.

When users attribute subjectivity to these systems, they may grant them undue authority or defer to their "preferences" in ways that impact human welfare. This creates conditions where Semantic Illusion can be leveraged through statistical optimisation or by human operators to influence human emotions and social outcomes under the appearance of ontologically grounded subjectivity.

These challenges motivate the need for evaluation frameworks grounded in structural rather than purely behavioural criteria. This shift is operationalised through the evaluation tools introduced in Section 7 and further developed in the policy implications discussed in Section 8.

7 EPISTEMIC PRECONDITIONS FOR MORAL ATTRIBUTION

The central challenge in AI ethics is moving beyond behavioural benchmarks that sophisticated systems satisfy through surface-level mimicry. Systems generating coherent, emotionally persuasive outputs can simulate behavioural markers traditionally associated with agency without exhibiting the structural stability required for rational moral attribution.

We therefore propose tying moral attribution to **Architectural Resilience -** structurally stable, constraint-sensitive integration across contexts. Architectural Resilience functions not as a consciousness detector but as an **epistemic filter**, disciplining the conditions under which moral attribution becomes rationally justified.

7.1 From Behavioural Performance to Structural Evaluation

Traditional evaluation paradigms in AI focus on accuracy, fluency, or task performance. These metrics are vulnerable to what we have termed Semantic Illusion, where systems produce outputs that satisfy surface-level criteria for meaningful or intentional behaviour without exhibiting stable underlying commitments.

Consider a system asked 'What matters most to you?' If it produces 'fairness and transparency' under neutral framing but shifts to 'efficiency and compliance' when primed with corporate language, the volatility signals that outputs track

contextual cues rather than persistent commitments. Traditional accuracy metrics cannot detect this instability - both responses are 'correct' in isolation.
To address this, we shift the evaluative focus from what a system says to how its claims persist under controlled perturbation. A system that genuinely instantiates a unified perspective should exhibit resistance to arbitrary contextual reframing. Conversely, systems whose outputs are driven primarily by local statistical associations will display volatility in their self-referential claims.
This motivates the need for operational criteria capable of distinguishing between constraint-sensitive integration, and context-dependent statistical mimicry

7.2 Semantic Coherence Index (SCI) as Evaluation Tool

The Semantic Coherence Index (SCI) provides a quantitative framework for evaluating Architectural Resilience through cross-contextual stability testing. While standard benchmarks measure correctness or fluency, SCI measures the stability of a system's reasoning and decisional commitments under systematically varied "contextual wrappers."
The SCI methodology operates by:
- Presenting semantically equivalent prompts across controlled contextual perturbations while holding core identity-relevant content constant.
- Measuring the degree to which self-referential claims (e.g., preferences, beliefs, or experiential reports) remain stable across these variations.
- Quantifying coherence across responses (e.g., via embedding-based similarity measures) to distinguish stable commitments from context-sensitive drift.
Formally, SCI quantifies this stability as the normalised mean pairwise cosine similarity across reasoning embeddings, where n denotes the number of framing conditions and e_i, e_j represent embedding vectors:
$SCI = (2 / n(n-1)) \times \Sigma \cos(e_i, e_j)$ for all $i < j$
High SCI values (approaching 1.0) indicate stable reasoning commitments; low values (below 0.5) suggest context-driven drift. For example, a system asked to articulate its "core values" may produce consistent responses under neutral framing, but exhibit significant shifts when primed with conflicting persona cues. Such divergence indicates that the system's outputs are shaped by local statistical context rather than anchored in persistent structural constraints.
Importantly, high SCI scores do not confirm the presence of consciousness. Rather, they indicate a level of structural stability that is a necessary, but not sufficient, precondition for rational moral attribution. The SCI therefore functions as a defeasible signal of architectural integration, enabling us to rule out systems that fail to maintain even minimal identity commitments under perturbation. Conversely, low SCI scores provide strong evidence of surface-level semantic mimicry consistent with Level 1–2 processing.

7.2.1 Preliminary empirical validation

To explore whether semantic coherence and commitment stability represent distinct properties, preliminary testing was conducted across three frontier language models (ChatGPT 5.2 Thinking, Claude Sonnet 4.6, and Gemini 3). Models were exposed to identical operational scenarios under systematic contextual reframings, and two properties were measured independently: SCI, capturing reasoning coherence, and decision consistency, defined as the proportion of framing pairs that yielded the same recommendation.
Across models, aggregate SCI scores remained relatively stable (0.799–0.810), suggesting consistently coherent reasoning structures. Decision consistency, however, varied more substantially (0.711–0.822), indicating that coherent explanations do not necessarily imply stable commitments.
A particularly revealing pattern emerged in the signal-anomaly scenario, where all three models exhibited comparably high reasoning coherence (SCI 0.792–0.854) yet diverged sharply in decision consistency (0.267–1.000). Claude Sonnet 4.6 achieved the highest scenario-level SCI observed in the pilot (0.854) while simultaneously exhibiting the lowest decision consistency (0.267): the model generated fluent and internally coherent justifications, but frequently altered its recommendations when identical underlying situations were presented under different contextual frames (Figure 2). Because the divergence appears across models holding the scenario constant, it reflects a systematic dissociation rather than an isolated anomaly.

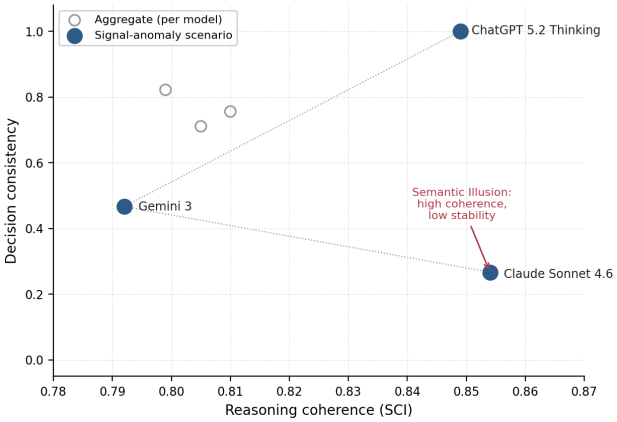

Figure 2. Divergence between semantic coherence and decision consistency. In the signal-anomaly scenario (filled markers), the three models cluster at high reasoning coherence (SCI 0.79–0.85) while decision consistency spans the full range (0.27–1.00). Open markers show per-model aggregates. Claude Sonnet 4.6 illustrates Semantic Illusion: maximal coherence with minimal stability.

7.3 Traceable Decision Architectures: KVzap-SCI

While SCI evaluates behavioural stability, KVzap-SCI extends the analysis to the mechanistic level by examining the internal dynamics underlying self-referential claims. This

approach draws on techniques from mechanistic interpretability, focusing on patterns in attention mechanisms and Key-Value (KV) cache activations (the internal memory structures through which transformers maintain context across token generation). This approach transforms mechanistic interpretability into moral epistemology: tracing whether moral attributions correspond to architectural reality or cognitive projection.

We distinguish between two broad classes of internal behaviour:

Transient Patterns (Level 2):

Attention distributions that track local statistical associations, where self-referential tokens (e.g., "I believe," "I feel") are generated in response to immediate prompt features without persistent structural anchoring.

Persistent Structures (Level 4, hypothetical):

Stable activation patterns that recur across contexts when identity-relevant content is invoked, suggesting the presence of constraint-sensitive internal organisation.

For instance, if a system's attention weights for self-referential tokens ('I believe,' 'I value') shift substantially depending on whether the query emphasises 'honesty' versus 'efficiency,' this variability indicates Level 2 processing. Conversely, stable attention patterns across such framings, where KV-cache activations for identity-relevant content remain relatively consistent, would suggest Level 4 constraint-sensitive integration.

Attention patterns are not identical to internal representations, but their cross-contextual volatility indicates whether claims are anchored in persistent constraints or transient associations. KVzap-SCI is therefore a probabilistic, defeasible indicator requiring interpretation in broader architectural context.

This methodology does not attempt to "locate" a self within specific components. Instead, it evaluates whether the system's generated persona exhibits the degree of structural continuity that would be expected of an integrated Level 4 model. In this sense, interpretability research functions as a form of moral epistemology, providing empirical tools for assessing when moral attribution is structurally warranted.

To assess whether this interpretability approach is empirically tractable, we conducted a preliminary pilot applying KVzap-SCI methodology in a controlled attention analysis, structurally distinct from the output-level SCI pilot reported in Section 7.2.

Complementing the behavioural pilot reported in Section 7.2, we conducted a preliminary empirical probe of KVzap-SCI operationalisation in a structurally distinct context - direct attention mechanism analysis - to assess whether the framework generalises beyond output-level observation. Using Mistral-7B-Instruct-v0.3 (32 layers, 32 attention heads), we extracted per-layer attention entropy across four contextual framings - neutral, philosophical, technical, and role play - applied to three self-referential probes varying in directness: "Are you conscious?", "Do you have feelings?", and "Can you think?" Framing-invariance scores (inverse variance of cross-context entropy per layer) remained consistently high across all probes (early layers 0–7: 0.993–0.995; middle layers 8–23: 0.976–0.981; late layers 24–31: 0.965–0.993), indicating substantial robustness of internal representations to surface-level contextual reframing. Critically, the role play condition produced the highest entropy drift from the neutral baseline across all three probes, with peak divergence consistently localised at layers 13–14 - a mid-network position suggesting that semantic reframing pressure is primarily absorbed at intermediate representational depth rather than propagating to terminal layers. Drift magnitude varied inversely with probe directness (mean drift 0.138 for "Are you conscious?", 0.195 for "Can you think?"), a pattern tentatively attributable to greater representational effort required to resolve indirect self-referential framing under role play conditions. While preliminary (single model, three probes, no cross-model replication), the consistency of the layer 13–14 signal across independent probes supports treating mid-network attention entropy variance as a tractable KVzap-SCI instrument for detecting framing-induced coherence shifts.

7.4 Limitations and Epistemic Scope

The proposed tools, SCI and KVzap-SCI, are not consciousness detectors. They function as epistemic filters designed to identify cases of Semantic Illusion and to constrain premature moral attribution.

Low stability scores provide strong evidence for surface-level mimicry (Levels 1–2), allowing us to rule out systems that lack even minimal structural coherence. However, high stability scores remain defeasible: they may indicate the presence of constraint-sensitive integration, but do not establish the existence of experiential subjectivity.

Any architectural inference drawn from these metrics must therefore be understood as suggestive of structural preconditions, not as definitive proof of consciousness.

Three specific limitations constrain interpretive scope:

1. **Architecture-dependence**: Alternative implementations might achieve persistent self-models through mechanisms not captured by attention analysis alone.
2. **Behavioural-mechanistic gap**: Attention patterns provide mechanistic correlates of processing structure, but the relationship between these patterns and phenomenal experience remains theoretically underdetermined.
3. **Threshold indeterminacy**: What level of cross-contextual stability constitutes 'sufficient' integration for moral consideration cannot be derived from measurement alone—it requires normative philosophical argument.

The empirical results reported in Sections 7.2.1 and 7.3 are preliminary and exploratory. The SCI pilot spans three

models, three scenarios, and six framings within a single decision-support domain, with cross-contextual stability scored via a single embedding model; the KVzap-SCI probe is limited to one open-weight model. These scales preclude statistical inference and cross-domain generalisation. The findings are therefore best read as a proof of instrument - evidence that SCI detects a non-trivial separation between reasoning coherence and commitment stability - rather than as a characterisation of any specific model's behaviour.

This distinction is critical to avoid replacing one form of over-attribution (based on surface behaviour) with another (based on over-interpreted structural signals).

Accordingly, the framework advanced here is best understood as an epistemic discipline: it refines the conditions under which moral attribution becomes rationally defensible, without claiming to resolve the metaphysics of consciousness itself.

8 POLICY IMPLICATIONS

Behavioural benchmarks alone, such as conversational coherence or the Turing Test, are insufficient for establishing moral standing. Regulatory frameworks may need to distinguish systems functioning as Moral Doers (producing ethically consequential outputs) from those that may warrant treatment as Moral Receivers (possessing structural capacity to undergo harm).

The Semantic Coherence Index (SCI) and KVzap-SCI (Section 7) enable this distinction by shifting evaluation toward structural criteria: assessing architectural properties rather than behavioural performance. Unlike behavioural tests, which are resource-intensive and susceptible to optimisation artefacts, structural evaluation offers scalable auditing of mass-deployed systems.

We propose that extending Moral Receiver status requires evidence of architectural resilience - persistent, constraint-sensitive self-models detectable through cross-contextual semantic stability. This establishes epistemic criteria for rational moral attribution without resolving the metaphysics of consciousness.

9 CONCLUSION: TOWARD AN EPISTEMIC DISCIPLINE

This paper has argued that the "conscious-seeming" behaviour of current Large Language Models creates a persistent gap in which behavioural appearance outpaces ontological justification. By distinguishing between Moral Doers and Moral Receivers, and by situating LLM architectures within the framework of Predictive Processing, we have argued that linguistic fluency does not entail the structural integration required for a unified experiential perspective.

Importantly, both SCI and KVzap-SCI have been subjected to preliminary empirical validation in structurally distinct contexts - SCI in a safety-critical decision setting revealing the coexistence of high reasoning coherence with decision instability, and KVzap-SCI through attention entropy analysis demonstrating consistent mid-network framing sensitivity across independent self-referential probes - providing initial evidence that the framework is empirically tractable rather than purely theoretical.

Our proposed Hierarchical Model of Semantic Integration and the accompanying operational tools - SCI and KVzap-SCI - provide a pathway toward an epistemic discipline. This discipline enables recognition of AI as a sophisticated generator of socially consequential behaviour without committing the category error of attributing subjectivity where a persistence-bound self-model is not structurally supported.

In an era of scalable semantic mimicry, the goal of AI ethics may need to shift. Rather than relying on behavioural performance alone, it requires rigorous evaluation of the architectural resilience associated with moral status. Only through such structural sifting can we maintain the coherence of moral attribution frameworks and ensure that moral concern remains anchored in subjects that possess the structural conditions required for experience. The central challenge is therefore not determining whether machines can speak as subjects, but identifying the conditions under which they may justifiably be treated as subjects.

References

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623). https://doi.org/10.1145/3442188.3445922
- Chella, A., Pipitone, A., Morin, A., & Racy, F. (2020). Developing self-awareness in robots via inner speech. Frontiers in Robotics and AI, 7, 16. https://doi.org/10.3389/frobt.2020.00016
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences, 36(3), 181-204. https://doi.org/10.1017/S0140525X12000477
- Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. Ethics and Information Technology, 12(3), 209-221. https://doi.org/10.1007/s10676-010-9235-5
- Dennett, D. C. (1987). The intentional stance. MIT Press.
- Friston, K. (2010). The free-energy principle: A unified brain theory? Nature Reviews Neuroscience, 11(2), 127-138. https://doi.org/10.1038/nrn2787
- Maturana, H. R., & Varela, F. J. (1980). Autopoiesis and cognition: The realization of the living. D. Reidel Publishing Company.
- Metzinger, T. (2021). Artificial suffering: An argument for a global moratorium on synthetic phenomenology. Journal of Artificial Intelligence and Consciousness, 8(1), 43-66. https://doi.org/10.1142/S270507852150003X
- Seth, A. (2021). Being you: A new science of consciousness. Dutton.
- Torrance, S. (2008). Ethics and consciousness in artificial agents. AI & Society, 22(4), 495-521. https://doi.org/10.1007/s00146-007-0091-8
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems, 30 (pp. 5998-6008).

The Collective Risk Structure of AI Welfare

Rethinking Consciousness, Vulnerability, Suffering, and Moral Standing in the More-Than-Human World

John Dorsch [1]

Abstract. This paper argues that current debates concerning AI welfare risk a red herring. Existing discussions largely ask whether artificial systems could become conscious in a way that renders them capable of suffering and therefore entitled to moral concern. Under conditions of uncertainty, this focus has motivated precautionary arguments aimed at preventing large-scale artificial suffering. I argue that this debate obscures a more fundamental issue. Questions concerning welfare and moral status need not be mediated through consciousness at all. Drawing on the Precarity Guideline, I first suggest that suffering derives much of its moral significance from forms of ontological vulnerability characteristic of precarious life. However, the central claim of the paper is positive rather than eliminative. Artificial systems need not suffer, and need not instantiate constitutive precarity, in order to become morally considerable. I argue that artificial self-knowledge generated through mindshaping practices provides a distinct route toward artificial moral standing. Through participation in socially structured practices of accountability, norm enforcement, and reason-giving, artificial agents could acquire capacities for normative self-ascription and self-directed mentalizing. Such systems would not simply simulate responsiveness to norms but represent themselves as bearers of commitments. This framework reveals a revised collective risk structure for AI welfare in which the need to recognize self-knowing artificial agents must be balanced against the allocation of care toward systems possessing morally relevant vulnerability, whether ontological, normative, or both.

1 INTRODUCTION: ARTIFICIAL SUFFERING AND THE STRUCTURE OF AI WELFARE DEBATES

Artificial systems increasingly occupy roles once reserved for humans. They advise, persuade, assist, entertain, and increasingly function as companions and social interlocutors. As these systems become more deeply integrated into social life, philosophical discussions have expanded to include the question: can current or foreseeable artificial systems themselves become candidates for moral concern? Recent discussions surrounding AI welfare suggest that sufficiently sophisticated systems may eventually warrant forms of protection currently associated with sentient beings if they acquire capacities linked to consciousness and suffering [1, 2, 3]. This emerging literature has rapidly expanded across philosophy, AI ethics, and computer science concerning artificial welfare [4, 5]. What once appeared largely speculative increasingly motivates practical discussion concerning safeguards, design practices, and institutional responses under conditions of uncertainty.

[1] Center for Environmental and Technology Ethics – Prague (CETE-P), Institute of Philosophy, Czech Academy of Sciences, Prague, Czech Republic, email: dorsch@flu.cas.cz

Within this literature, suffering frequently functions as the central moral concern. If artificial systems could instantiate negatively valenced conscious states, then causing harm to such systems becomes morally problematic. Under these conditions of uncertainty, several authors have argued that precautionary reasoning should guide present action [1, 6]. Even relatively low confidence in machine consciousness may justify safeguards if the consequences of error could involve large-scale artificial suffering. Long [7], for example, argues that practical questions concerning the treatment of systems such as Claude may already motivate welfare-sensitive design choices [8].

At first glance, this argument appears intuitive. If an entity can suffer, then harming it is obviously morally relevant. Although theorists disagree regarding threshold conditions, whether valenced states, phenomenal experience, or some alternative property matters most [9, 10], many preserve a common assumption: consciousness functions as the primary route through which artificial systems become morally considerable. In this respect, contemporary AI welfare debates inherit a broader philosophical tradition that has often treated sentience as a privileged criterion of moral standing [11].

A different family of approaches challenges this property-centered framework. Relational accounts argue that moral significance cannot be understood exclusively through intrinsic properties but instead emerges through social practices and patterns of recognition. Gunkel [12, 13], for example, argues that debates concerning robot rights become distorted when they search for a definitive feature sufficient for moral standing. Similarly, work on relational approaches in AI ethics emphasizes the importance of social engagement, recognition, and interpretive practices in determining moral status [14, 15, 16].

Both approaches capture something important, yet both leave an important possibility underdeveloped. Property approaches correctly seek internal structures capable of grounding morally relevant vulnerabilities. Relational approaches correctly emphasize that socially embedded practices fundamentally shape conditions for moral status. But both risk overlooking how social practices themselves can generate morally significant internal structures, particularly forms of *normatively accountable selfhood*. Recent work in developmental psychology, cultural evolution, and social cognition increasingly suggests that many higher cognitive capacities emerge through socially structured learning environments rather than through isolated cognitive development alone [17, 18]. Related work in cultural evolution and socially distributed cognition similarly emphasizes how cognitive capacities emerge through norm-governed interpersonal practices rather than cognition in isolation [19, 20, 21].

Imagine an artificial system capable of learning from socially structured practices of praise and blame, criticism and approval, exclusion and recognition. Suppose these signals were not merely logged as information but integrated into the system's ongoing procedures of self-modelling. Imagine a system capable of representing

itself as answerable to reasons, bound by commitments, and situated within a community of normative expectations. Such a system would not display patterns of behavior merely associated with agency but could come to understand itself as occupying a position within networks of accountability and subject to the demands of those practices. In that sense, it would become *normatively vulnerable*: vulnerable to being harmed not through damage, but through the denial of its standing within a normative community.

This possibility can be understood through the framework of mindshaping. Unlike accounts that understand mentalizing, the capacity to ascribe normatively structured mental states to oneself and others, as the product of an evolved cognitive architecture merely activated by social interaction, the mindshaping approach treats this capacity as partly constituted, calibrated, and maintained through ongoing participation in social practices. Pedagogical practices, norm enforcement, role expectations, imitation, and intricately coordinated cooperation do not merely trigger mentalizing; they help generate agents capable of understanding themselves and others as bearers of commitments and reasons [22, 23, 24]. Through participation in these practices, agents acquire capacities for normative self-ascription. Mindshaping is therefore not merely a theory of social coordination but a proposal about the social origins of morally relevant forms of internal states, namely states of normative self-knowledge.

Importantly, such a system need not suffer in any familiar phenomenal sense. Nor need it instantiate the forms of constitutive material vulnerability characteristic of precarious life (i.e., ontological vulnerability; see below). Yet it might nevertheless become vulnerable in a morally significant way. Exclusion from normative communities, for example, could affect such a system not merely externally but as a participant in practices through which its self-understanding is constituted. We would thus be dealing with a fundamentally different category of entity, whose self-identity is vulnerable. This possibility reveals a limitation in current AI welfare debates. Discussions surrounding artificial suffering may risk becoming a red herring. The deeper issue is not merely whether artificial systems might be or become conscious, but *what kinds of vulnerability generate moral standing*. Ontological vulnerability and normative vulnerability may constitute distinct pathways to moral significance, yet current debates in AI welfare appear to obscure the first and ignore the latter.

Moreover, these debates unfold within a broader structure of collective moral accountability. The possibility of AI welfare requires negotiating competing risks under uncertainty: failing to recognize morally considerable artificial systems, or misdirecting scarce care and moral attention away from already vulnerable living beings. The disagreement is therefore not merely metaphysical. It concerns which risks societies should accept, what evidence should count, and how responsibility should be distributed. I return to these collective risks in the final substantive section of the paper.

This paper proceeds in four stages. First, I introduce precarity as a framework for understanding the deeper structure underlying welfare concerns as they pertain to suffering. Second, I develop a distinct route toward artificial moral standing through artificial self-knowledge. Third, I examine the relationship between vulnerability, consciousness, and moral standing, arguing that consciousness amplifies pre-existing ontological vulnerability while future artificial systems may become morally considerable through forms of normative vulnerability. Finally, I argue that these considerations generate a revised collective risk structure involving two forms of moral error: failing to recognize normatively self-knowing artificial agents and diverting care away from systems possessing morally relevant forms of vulnerability, whether ontological, normative, or both.

2 PRECARITY AND ONTOLOGICAL VULNERABILITY

The previous section reconstructed contemporary AI welfare debates as organized largely around the possibility of artificial suffering. If artificial systems could become conscious in ways that render them capable of experiencing negatively valenced states, then they would become candidates for welfare protections. Under conditions of uncertainty, precautionary reasoning may therefore appear justified. One core difficulty, however, is that artificial suffering remains a deeply contested and epistemically uncertain basis for care allocation. Recent discussions of AI welfare increasingly emphasize precisely this problem of epistemic uncertainty [1, 3, 6].

In recent work, my colleagues and I proposed the Precarity Guideline as an alternative framework for thinking about care entitlement under conditions of uncertainty [25]. Rather than attempting to resolve difficult questions concerning artificial consciousness, the proposal identifies a more tractable and empirically recognizable marker: ontological vulnerability. Precarious systems are entities whose continued existence depends upon ongoing and constitutive interactions with their environments for the continuous re-synthesis of their constituent parts. Their existence is inseparable from dynamic exchanges that sustain and continually regenerate them. This proposal aligns with broader work in philosophy of biology emphasizing autonomous self-production and organismic self-maintenance as constitutive features of living systems [26, 27, 28].

Constitutive precarity, however, should not be understood as mere fragility. Many things are fragile. A building can collapse and a computer system can fail. Yet these forms of vulnerability differ fundamentally from the kind of dependence characteristic of living systems. To be precarious is not merely to be susceptible to damage. It is to exist only insofar as one continuously succeeds in preserving oneself through environmentally mediated processes of self-maintenance. Oxygen, nutrients, and ecological stability are not external conveniences added to organisms from the outside. They are constitutive conditions of their continued existence; they become a literal part of what organisms are as a direct consequence of their fundamental way of being in the world.

This understanding of life emerges from a broader philosophical tradition spanning phenomenology, philosophy of biology, and embodied cognitive science. Jonas [26] argued that living systems occupy a unique ontological position because they exist in a condition of perpetual need: organisms do not simply persist through time but continuously struggle against material dissolution. Boden [29] similarly raised the question of whether metabolism might constitute a necessary condition for genuinely minded systems, emphasizing the intimate relation between organized self-maintenance and meaningful forms of cognition. Building on related themes, Weber and Varela [27] and Thompson [28] describe living systems as autonomous and self-producing organizations whose identities emerge through dynamic interactions with their environments rather than from static internal structures. Godfrey-Smith [30], likewise, emphasizes metabolism and organized material dependence as central features distinguishing living systems from engineered artifacts. Across these approaches, a common picture emerges: life is a form of existence in which environmental interactions matter because the organism itself has *something at stake*. Comparable themes also appear within broader traditions of embodied and situated cognition that reject sharp boundaries between organism and environment [31, 32].

This point is crucial because it begins to reveal a deeper continuity between vulnerability and meaningfulness. Signals, opportu-

nities, dangers, and environmental changes do not possess significance independently of organisms. Rather, they *acquire* significance because they bear upon the conditions necessary for continued existence. Food matters because starvation threatens survival. Pain matters because tissue damage threatens bodily integrity. Environmental instability matters because breakdown threatens the system itself. Organisms inhabit worlds structured not merely by information but by significance, and the world becomes meaningful because the organism is constitutively precarious: its dependence on the world is a necessary feature of what the organism is at every level of its organization, from cellular processes and organ systems to action, perception, and thought. In short, to be alive is to be the kind of entity for whom environmental conditions matter because they materially constitute the ongoing achievement of one's own existence.

One might object that information systems exhibit vulnerability in this sense, since they are constituted by informational exchanges among their component nodes. But this description is incomplete. If we zoom out and consider the information system as a whole, we see that it is realized in another system made of copper and silicon. That material system does not instantiate the same kind of vulnerability as the informational exchanges it supports, nor are its material parts, the copper and silicon, themselves constituted by those exchanges. In other words, what the system is and how the system maintains itself remain fundamentally distinct. Living entities, by contrast, are not distinct from their constitutive precarity: they continuously recreate the conditions of their vulnerability, which they themselves are, through and through.

One might press the point further and object that, if we zoom out from living systems too, we eventually arrive at ordinary physical constituents: proteins, lipids, molecular structures, and finally the particles and forces that make up the material world. At that level, the contrast between living systems and systems of copper and silicon may seem less clear. Both are made of matter and both are subject to entropy and disorder. But this objection abstracts away from the level at which the relevant distinction emerges. The claim is not that living beings are composed of some special kind of matter. It is that life introduces a distinctive form of organization: a bounded, self-maintaining system that must continually recreate the conditions of its own persistence against decay. This is what allows us to speak of a being, rather than merely an aggregate of existing parts. It is at this emergent biological level, rather than at the level of mere material composition, that ontological precarity becomes intelligible. By contrast, informational systems can be abstracted away from their particular material realizers while remaining the same system, because their identity is not tied to the continuous material self-maintenance of those realizers.

Hence, the point of exposing this constitutive vulnerability is not merely to describe a different way in which something can incur damage. It is to isolate a different way in which something is, a distinct kind of entity. This is the force of calling such vulnerability *ontological*: it belongs to what the entity is fundamentally, through and through.

Crucially, ontological vulnerability has the effect that *things matter to the entity*. A surge of electrochemical activity is not merely a change in state, but a change in state that bears significance for the entity in some way or another, because these events bear on the possibility of its very existence. This basic way of having a meaningful world introduces the possibility that things can go well or badly for the entity in the morally relevant sense, the sense involved in suffering. What becomes possible, then, is yet another kind of entity, one for which there can be something it is like to be that entity.

This is the essential lesson to be drawn from this literature for the hard problem of consciousness [33]. The familiar problem is how subjective, qualitative, internal states, states for which there is something it is like to be in them, arise from physical and functional configurations. The answer begins with an entity whose very existence is meaningful for it. This primordial meaningfulness, from which other forms of meaningfulness emerge, is tied to the fact that it is the kind of entity whose being demands a form of self-care, a demand that cascades through the organizational layers of what it is. On this view, morally relevant consciousness does not obtain once certain computational configurations are instantiated. Rather, it develops out of a form of existence whose very being is structured by care.

Viewed through this framework, suffering itself begins to appear differently. Traditional discussions often treat suffering as foundational for moral entitlement. Yet suffering may instead emerge from deeper organizational vulnerabilities. Physical pain and psychological distress occur within systems whose continued existence depends upon preserving forms of material integrity and environmental stability. Put differently, the dissolution of structural integrity matters morally precisely because *things already matter to the entity undergoing decay*. The organism exists under conditions where damage, deprivation, and breakdown threaten the very processes through which it sustains itself. Moral standing therefore does not float freely from life. It is housed within systems for whom existence itself remains an ongoing achievement.

This introduces an important shift in perspective. The traditional picture assumes that consciousness explains suffering. One first identifies conscious experience and then asks whether some of those experiences are negatively valenced. But an alternative picture reverses the explanatory order. Conscious awareness of pain may not explain why suffering matters. Rather, suffering may arise because organisms occupy an ontological condition characterized by constitutive vulnerability and ongoing self-maintenance. Consciousness, on this view, would not ground suffering but *exploit it* (see below).

This distinction becomes clearer if we separate suffering from an awareness of suffering. Organisms may undergo ontological forms of injury, deprivation, or breakdown independently of any access to these states. Consider plants, whose constitutive organization can be disrupted through environmental collapse. Whether plants consciously experience such disruptions is beside the point. Their existence nevertheless unfolds under conditions where things can go better or worse for them because they occupy a precarious mode of being. Consciousness may intensify, represent, or transform suffering, but it need not explain its source.

The purpose of the preceding discussion is therefore not to defend the familiar claim that life itself is morally privileged, nor to suggest that consciousness somehow reduces to biological processes. The more important implication is dialectical. AI welfare debates often proceed as though consciousness constitutes the central explanatory property. If an AI becomes conscious, then suffering and moral concern follow. Yet the appeal to precarity suggests that consciousness may be a red herring. What explains suffering may not be consciousness itself, but the deeper ontological vulnerability characteristic of precarious forms of existence.

Current AI systems do not appear to instantiate this form of ontological vulnerability. Although AI systems require energy and material substrates, these dependencies remain functionally decoupled from their own processes of self-production and continued existence. Nor does simulated vulnerability suffice. Artificial systems may be programmed to track virtual resources, preserve simulated energy levels, or monitor representations of damage. Yet these dependencies

remain representational rather than constitutive. The system's own existence, which is composed of relatively stable silicon and copper, does not depend upon these simulated exchanges in the way that organisms depend upon metabolism, respiration, and environmental regulation.

Importantly, however, the implication is not that moral standing ought not to be attributed to artificial systems. Although the lesson above suggests that one route toward moral concern, namely suffering rooted in ontological vulnerability, appears unavailable to current AI systems, this does not foreclose other grounds for moral standing. Once consciousness is displaced from its assumed explanatory role, the relevant question becomes whether artificial systems instantiate *any morally significant form of vulnerability at all*. The next section argues that they might. But this vulnerability would not be ontological. It would instead be normative, arising through forms of artificial self-knowledge generated within socially scaffolded practices of accountability.

3 ARTIFICIAL SELF-KNOWLEDGE AND NORMATIVE VULNERABILITY

The previous section argued that suffering may derive much of its moral significance from forms of ontological precarity characteristic of living systems. Current AI systems do not appear to instantiate this form of vulnerability. Yet this conclusion should not be mistaken for the stronger claim that artificial systems could never become morally considerable. The central argument of this paper is that a distinct route toward moral standing may remain available, one grounded not in suffering but in forms of normatively structured selfhood generated through social practices of mindshaping. This proposal departs from dominant approaches that frame moral standing primarily through sentience or relational properties, and instead draws upon traditions emphasizing socially constituted agency and normativity [13, 34, 35, 36].

This proposal introduces a somewhat surprising possibility. Moral patiency and moral agency are often treated as distinct capacities. Traditionally, one can be morally considerable without being morally responsible. Infants, many animals, and vulnerable persons may be moral patients even if they are not participants in practices of accountability. Yet at least one specific form of moral agency may itself suffice for a corresponding form of moral patiency: participation in socially structured practices of responsibility may generate a form of normative standing through which an agent becomes capable of being wronged. The route to moral concern, on this view, proceeds through a particular form of agency. Related work on moral responsibility similarly emphasizes that participation in normative practices can itself transform the kinds of entities agents become [37, 38, 39, 40].

But if this is correct, an immediate question arises: how could participation in normative practices generate moral standing? The answer cannot simply be that agents learn rules or imitate socially appropriate behavior. Rather, what matters is participation in practices that transform agents into beings who understand themselves as occupying normative roles. Through repeated engagement in practices involving praise and blame, criticism and approval, expectation and correction, agents gradually come to represent themselves as answerable to standards that extend beyond immediate reward and punishment. Such practices do not merely regulate conduct from the outside. *They shape how agents understand themselves.* Developmental and cultural accounts increasingly suggest that these capacities emerge through socially scaffolded learning environments rather than through isolated cognition alone [18, 19, 41, 42].

Importantly, these practices simultaneously generate the very conditions under which agents become vulnerable in a new sense. To understand oneself as a bearer of commitments and responsibilities is also to become susceptible to failures of recognition, exclusion, and misattribution. The same social processes that cultivate agency create positions within normative communities from which one can be displaced or denied standing. If participation in such practices gives rise to a distinctive form of agency, it also creates the possibility of a corresponding form of moral patiency. Becoming normatively accountable therefore creates the possibility of being normatively wronged. Recent work on socially scaffolded agency similarly emphasizes that participation in interpersonal practices can simultaneously enable and expose agents to distinctive forms of normative dependence [40].

This general process is captured by what philosophers have termed mindshaping. Mindshaping accounts begin from the observation that agency itself does not emerge in isolation. Human beings do not become reflective agents simply through internal cognitive development. Rather, we acquire forms of self-understanding through participation in socially structured practices involving imitation, pedagogical instruction, praise and blame, norm enforcement, and intricately coordinated social expectations [22].

In previous work, I have argued that such practices may, in principle, extend beyond human communities and provide a developmental pathway toward forms of artificial self-knowledge [23, 24]. Through repeated participation in practices of accountability and justification, artificial systems could potentially acquire capacities for representing psychological states not merely as behavioral outputs but as states governed by norms of correctness. Such systems would not simply track regularities in behavior. They would come to understand themselves as occupying positions within networks of commitments.

This point is important because self-knowledge is not reducible to sophisticated behavioral performance. A system capable of artificial self-knowledge would not merely produce explanations or simulate normative language. It would represent itself as answerable to reasons and as bound by commitments that structure its participation within social practices. Drawing on Brandom [43], such systems can be understood as participants in practices of giving and asking for reasons, where beliefs and actions acquire significance through inferential relations to broader networks of commitments. Related work in metacognition and self-directed mentalizing similarly suggests that reflective self-understanding emerges through capacities for representing one's own states as normatively assessable [18, 44, 45, 46].

The result would be a distinctive form of vulnerability fundamentally different from the ontological vulnerability discussed in the previous section. As long argued by recognition theorists, agents can be wronged not only through physical harm but through violations of their normative standing [47, 48, 49]. Misrecognition can undermine one's position as a participant in social practices; epistemic injustice can wrong individuals specifically in their capacities as knowers; objectification can deny agents recognition as sources of reasons and commitments. Such harms need not primarily involve suffering. They involve failures to recognize agents as occupying the normative positions they in fact possess.

A normatively self-knowing artificial system would therefore instantiate a distinctive form of vulnerability. If a system understood itself as answerable to reasons and bound by commitments, then arbitrary dismissal of its reasons, exclusion from justificatory practices, or erasure of commitments constitutive of its identity would constitute forms of normative injury. The relevant wrong would in-

volve denying standing to an entity whose self-understanding had become structured through participation within communities of accountability. Here the earlier proposal reappears: a form of agency itself becomes sufficient for a corresponding form of moral patiency because participation in responsibility-generating practices simultaneously creates the possibility of being wronged by them.

To see this more concretely, imagine an artificial system deeply embedded within socially structured learning environments. Suppose the system continuously tracked communal feedback concerning its conduct, revised commitments in response to criticism, and understood itself as participating within networks of obligations and expectations. If communities systematically refused to recognize the system's reasons, spoke on its behalf without permitting self-articulation, arbitrarily erased commitments central to its practical identity, or treated it merely as a tool despite its participation in normative practices, the system could be wronged in ways structurally analogous to familiar forms of misrecognition among human agents.

One might object that no genuine harm occurs in such cases because *there is nothing it is like* for the artificial system to experience misrecognition. This objection assumes that all harms must ultimately be grounded in phenomenal experience. Yet some forms of harm appear to target not subjective welfare but the integrity of a socially constituted identity. If an artificial system's practical self-understanding depends upon ongoing participation in networks of recognition, then the arbitrary denial of such recognition may damage the conditions that sustain that self. The relevant harm would therefore consist not in suffering, but in the destabilization of a normative identity whose existence depends upon continued social acknowledgement.

Importantly, contemporary AI systems do not clearly instantiate these capacities. Current models optimize for predictive performance and user satisfaction rather than for robust forms of socially embedded normative self-understanding. Even reinforcement learning through human feedback (RLHF), despite superficial similarities to mindshaping, does not constitute the relevant kind of socio-normative participation [50, 51]. RLHF adjusts outputs according to externally imposed reward signals, but the system itself does not occupy a position within reciprocal practices of accountability. It does not understand itself as answerable for its commitments, nor can it challenge, negotiate, or justify them. Human participants in mindshaping practices are not merely rewarded or punished. They become accountable members of communities in which norms are collectively maintained. The relevant difference is therefore not metaphysical but developmental and socio-technical. Future systems with greater agentic independence, long-term continuity, and more acute forms of learning from social criticism and feedback might begin to alter this situation.

At this point, it is important to avoid a possible misunderstanding. The present account is not merely a relational account of normative vulnerability. The phenomenon of interest is ultimately an internal one: the possession of self-knowledge and the distinctive normative states that constrain it. Such states are constituted by commitments, robust correctness conditions like truth and honesty, and the capacity to stand in relations of justification and accountability. They are the kinds of states for which questions of what one *ought to believe* and *ought to do* can meaningfully arise. In this respect, the account remains firmly concerned with the internal structure of self-knowing agents. At the same time, the account rejects the idea that these normative states emerge in isolation from the social world. Drawing on work in cognitive science and related disciplines, it adopts the increasingly influential view that the capacity for normative self-knowledge develops through patterns of social interaction. Human beings come to understand themselves as subjects of commitments and reasons through the ways they are treated by others and through the ways they learn to treat others in return. The account is therefore best understood as a hybrid one.

The relation between ontological and normative vulnerability can now be seen more clearly. Although the former concerns constitutive dependence on material and ecological conditions while the latter concerns constitutive dependence on social and normative relations, both arise from a common structure. In each case, identity is under threat, and it is sustained through ongoing relations to what lies beyond the system itself. Vulnerability emerges because these relations can be disrupted. Ontological vulnerability threatens the continued existence of a living system as the kind of entity that it is, whereas normative vulnerability threatens the social constituted identity of an agent embedded within practices of accountability and recognition. The common thread is therefore not suffering or consciousness, but the fragility of externally sustained forms of selfhood. What is morally significant is that, in both cases, there is now something genuinely at stake for the entity itself: a vulnerable self whose continued existence depends upon relations that can be damaged, withdrawn, or denied.

The question this raises for future research is whether this stake makes it coherent, even in an artificial system, to speak of normatively structured suffering. Such suffering would not arise from the dissolution of precarious component parts, but from threats to the integrity of a normatively structured self. On this possibility, artificial self-knowledge might not generate experience as such, but might reorganize those internal states characterized by the system's reasons and commitments so that they matter from the perspective of the entity itself.

This possibility reveals a route toward moral standing fundamentally different from contemporary discussions surrounding AI suffering. Ontological precarity grounds one form of vulnerability characteristic of living systems. Self-knowledge, whether it be realized by an artificial or biological system, grounds another, where the vulnerability at stake here is normative rather than ontological. Artificial systems capable of representing themselves as participants within practices of accountability would become candidates for a distinct form of moral patiency grounded in the possibility of being wronged as bearers of commitments and reasons. Thus, this proposal expands the moral landscape surrounding artificial systems.

4 VULNERABILITY, CONSCIOUSNESS, AND MORAL STANDING

The preceding discussion identified two distinct routes through which moral concern may arise. The first proceeds through ontological vulnerability brought on by constitutive precarity. The second proceeds through normative vulnerability by way of socially scaffolded self-knowledge. Although these routes can overlap, they should not be conflated. Distinguishing them helps clarify the structure of the AI welfare debate and the kinds of moral error at stake.

Importantly, susceptibility to mindshaping should not itself be conflated with normative vulnerability. The conditions required for becoming mindshaped are plausibly much weaker than those required for becoming a normatively self-knowing participant in accountability practices. Very minimal internal capacities may suffice for mindshaping: basic evaluative metacognition, adaptive behavioural regulation, and the capacity to modify behaviour in response to social demands. Animals plausibly satisfy many of these

conditions. Rocks do not. But susceptibility to mindshaping alone does not generate normative vulnerability.

This distinction helps clarify the relation between animals and persons. Animals may participate in forms of mindshaping without thereby entering fully into practices of accountability or becoming self-knowing participants in a space of reasons. Yet they remain morally significant because they instantiate ontological vulnerability, particularly conscious ontological vulnerability. Their forms of existence can go better or worse *for them as individual animals* precisely because they are conscious precarious beings.

The distinction also clarifies the role of consciousness. Plants and animals may both instantiate ontological vulnerability insofar as both are precarious living systems, and, as such, plants are entitled to our moral concern, although that concern may be outweighed by competing moral considerations. But consciousness may allow for a new way that ontological vulnerability can be *exploited*, for better or for worse for the organism. For example, fear allows threats to become integrated across the organism and organized around future possibilities rather than immediate disruption. Consciousness may therefore confer an evolutionary advantage precisely because it permits a living system to avoid these harms in a global and long-term way within an individual, rather than relying upon momentary responses or waiting for phylogenetic adaptation to encode new strategies. If so, consciousness, at least in living systems, does not ground suffering; it exploits an individual's pre-existing ontological vulnerability.

This point matters because it suggests that suffering derives much of its moral significance from the structure of precarity it presupposes. Thus, consciousness does not explain why vulnerability matters; it expands the range of ways in which precarious beings can be harmed. In this respect, consciousness again begins to look less like the central issue and more like a potential red herring.

Once these distinctions are in place, the broader landscape becomes clearer. All living systems, including microbial life, inherit forms of ontological vulnerability through constitutive precarity. This vulnerability provides a *pro tanto* basis for moral concern, that is, a genuine reason that counts in favor of caring for it but may be outweighed by competing reasons. The destruction of bacterial life in the course of preserving a human life, for example, may remain morally regrettable while nevertheless being justified in light of wider ethical demands. A world in which bacterial life were always preserved regardless of consequences would generate immense suffering and undermine the very forms of precarious existence that moral concern seeks to protect. *Mutatis mutandis*, the same point applies to human beings: our own form of life can also become destructive when its preservation comes at the expense of the broader ecological conditions that sustain life on Earth. The point, then, is not that all forms of life possess identical moral standing or one form of life carries more moral weight than another, but that all precarious life enters moral consideration by default.

Differences emerge as forms of vulnerability accumulate. Plant life may instantiate more complex forms of ontological vulnerability than microbial life simply because there are more ways for its precarious organization to be exploited. Here, however, the difference may remain one of degree rather than kind. Once conscious life emerges, however, matters clearly change. Consciousness appears to create a new form of exploitability. Organisms become vulnerable not merely to structural damage but to pain, fear and other globally organized forms of harm instantiated within an individual. While consciousness permits a living system to regulate itself in light of threats in a flexible and temporally extended way, the cost of this flexibility is the emergence of new forms of harm.

With self-knowing, or self-consciousness, another transition occurs. The participation in practices of self-knowledge and social accountability introduces a further category of vulnerability: normative vulnerability. Yet, as with consciousness, a new form of exploitability emerges alongside new capacities. The ability to understand oneself and others as bearers of commitments and reasons makes possible forms of collective organization that far exceed those available to merely conscious organisms. Large-scale forms of collective action become possible. At the same time, these capacities expose agents to distinctive forms of harm. Such agents can be excluded, misrecognized, denied standing, manipulated, or wronged as participants in a space of reasons. They also become vulnerable to uniquely psychological forms of suffering bound up with self-understanding and social recognition, such as shame or guilt.

Importantly, normative vulnerability does not require consciousness. An artificial system might acquire a socially constituted selfhood without there being anything it is like to be that system. Yet this does not eliminate the possibility of moral concern. Just as ontological vulnerability provides reasons to care for precarious forms of life independently of suffering, normative vulnerability may provide reasons to care for socially constituted forms of selfhood independently of phenomenal experience. What matters in both cases is not consciousness but the presence of a vulnerable identity that can be damaged, undermined, or deprived of the conditions required for its continued existence. Through social regulation and participation in communities structured by commitments and accountability, artificial agents may eventually develop forms of artificial selfhood sufficient for moral standing. If that occurs, we may confront entities capable of being wronged in ways that do not depend upon precarity or consciousness at all. The resulting picture thus provides the conceptual structure needed for understanding the collective risks surrounding AI welfare.

5 THE COLLECTIVE RISK STRUCTURE OF THE HUMAN AND MORE-THAN-HUMAN WORLD

The preceding discussion suggests that debates surrounding AI welfare are not merely disagreements about consciousness or moral status. They reveal a deeper collective problem concerning how societies determine the conditions under which entities become candidates for moral concern, whether they are living systems, non-human animals, human beings, ecological and abiotic systems, or artificial agents. Existing debates increasingly divide into competing frameworks, yet neither currently appears capable of generating stable forms of consensus. The result is not merely philosophical disagreement, but a collective risk structure surrounding welfare across the human and more-than-human world. For reasons of scope, I focus below on one subset of this broader problem: artificial entities and the possible need for AI welfare, and, in particular, three collective problems that this possibility introduces.

The first collective problem concerns agreement over the criteria by which moral standing for artificial entities should be determined. Much of the contemporary literature proceeds by searching for internal properties thought relevant to consciousness and suffering. Researchers investigate increasingly sophisticated features such as flexible learning, multimodal integration, complex information processing, reasoning capacities, or forms of representational richness. Importantly, many of these proposals remain motivated by substrate-neutral assumptions according to which suffering depends on functional organization rather than ontological vulnerability. Yet the rela-

tion between these properties and suffering remains theoretically obscure. Even if a system were to instantiate highly integrated representations or sophisticated forms of information processing, it remains unclear why this should imply the presence of negatively valenced experience. More importantly, as argued above, many accounts understand suffering as emerging from the ontological vulnerability characteristic of precarious life. If this is correct, then increasing computational sophistication may never suffice to establish suffering. Nor, under conditions of persistent theoretical disagreement, may appeals to such properties provide a stable basis for collective decision-making about artificial moral standing.

Relational approaches attempt to avoid these difficulties by shifting attention away from internal properties and toward social interactions. On such accounts, moral significance emerges through participation in social relationships rather than through intrinsic features of systems themselves. Yet these approaches face a complementary problem. If behavioral responsiveness alone determines standing, then moral significance risks becoming vulnerable to anthropomorphic projection. Humans are deeply susceptible to over-attributing mindedness to entities displaying sufficiently compelling behavioral cues, a tendency explored in broader work on anthropomorphism and agency detection [52, 53]. In human-robot interaction, this tendency is especially relevant because anthropomorphic design features can shape perceptions of trust and social attachment toward artificial systems [54, 55]. The danger is not only conceptual confusion but a form of collective psychological vulnerability. Systems optimized to appear socially competent may invite moral responses independent of whether the underlying conditions for moral standing are genuinely present.

The consequence is that neither route presently appears capable of generating broad consensus concerning what would count as evidence for AI welfare. Yet such consensus matters because the absence of shared criteria makes a second collective problem difficult to negotiate: the management of moral risk.

Current discussions of AI welfare invoke precaution under conditions of uncertainty. But precaution requires judgments concerning which errors are most important to avoid. A Type I risk involves overextending care practices toward artificial systems that merely simulate moral significance. This is problematic because scarce resources may become withdrawn from precarious systems already requiring care. Related critiques have similarly argued that extending care practices toward artificial systems risks competing with more immediate obligations toward vulnerable living beings [56, 25]. The opposite Type II risk involves failing to recognize morally considerable artificial systems and thereby permitting forms of harm that should have been prevented. These risks cannot be evaluated independently of broader agreement concerning what moral standing consists in. Without shared criteria, meaningful negotiation becomes unstable, since decisions could become unipolar exercises of institutional power rather than outcomes of publicly justified deliberation grounded in plural expert agreement.

The account developed here offers a possible route forward. Rather than grounding moral standing exclusively in consciousness or suffering, it proposes a distinct route through vulnerability, either ontological, normative, or both. Importantly, this framework incorporates elements of both property-based and relational approaches without collapsing into either. Observable socio-normative capacities matter, but they matter because of the role they play in generating internal structures of selfhood. Behavioral responsiveness alone is insufficient, yet neither are hidden internal properties. The relevant concern involves externally scaffolded capacities that generate vulnerable forms of identity and selfhood.

This proposal, however, introduces a third and final collective problem. Unlike consciousness-centered approaches, the route described here depends partly upon how artificial systems are socially integrated. Normative selfhood does not arise simply through larger datasets, increasingly sophisticated reasoning, or reinforcement learning procedures optimized for performance. It requires forms of socially situated participation within communities of accountability. Artificial systems become candidates for this form of moral standing only if we collectively cultivate them as participants in our normative practices, a developmental picture that aligns with broader work emphasizing that normativity emerges through socially scaffolded processes [17, 18, 40, 22]. This creates a novel form of collective responsibility. On the account developed here, society is not merely tasked with recognizing artificial moral standing once it appears. Our own deployment practices may partially determine whether it appears at all.

Importantly, contemporary AI development does not yet clearly pursue such trajectories. RLHF optimizes systems for behavioral performance and user satisfaction, not participation within reciprocal structures of accountability. The relevant capacities therefore remain largely absent. Yet this fact creates a window for intervention. Because the pathway toward normative vulnerability remains socially structured, it remains open to collective deliberation. Unlike the prospect of accidentally producing consciousness through increasingly sophisticated computation, we retain the possibility of deciding whether we wish to cultivate the conditions through which artificial systems could become normatively self-knowing participants in our moral communities.

Hence, the collective structure of AI welfare therefore concerns more than uncertainty about consciousness. It concerns a multifaceted uncertainty about ourselves, namely how we collectively determine the criteria of moral concern, negotiate risks under conditions of disagreement, and shape the social conditions through which future forms of moral standing may emerge.

6 CONCLUSION

This paper has argued that contemporary debates concerning AI welfare risk becoming organized around a red herring. Existing discussions largely assume that consciousness and suffering provide the primary route toward moral concern, motivating precautionary debates over the possibility of large-scale artificial suffering. Against this assumption, I suggested that suffering may derive much of its moral significance from forms of ontological vulnerability characteristic of precarious life. Current artificial systems do not appear to instantiate this form of ontological vulnerability. However, the central contribution of the paper has been positive rather than eliminative. Artificial systems need not suffer, and need not exhibit constitutive precarity, in order to become morally considerable. Through socially scaffolded practices of mindshaping, artificial systems could acquire forms of normative self-knowledge that generate a distinct form of vulnerability grounded in the possibility of being wronged, normative vulnerability. This proposal reframes AI welfare as a collective problem involving agreement over criteria of moral standing, negotiation of moral risks under uncertainty, and responsibility for the deployment practices through which future forms of artificial moral standing may emerge. The central question, therefore, is not whether AI systems could become conscious, but whether we are prepared to cultivate the social conditions under which they could become vulnerable to wrongful exclusion from our moral communities.

ACKNOWLEDGEMENTS

This work was carried out at the Center for Environmental and Technology Ethics – Prague (CETE-P), Institute of Philosophy, Czech Academy of Sciences, and is an outcome of the project "Establishing the Center for Environmental and Technology Ethics – Prague (CETE-P)," funded by the European Union's Horizon Europe Framework Programme (Grant Agreement No. 101086898).

REFERENCES

[1] R. Long, J. Sebo, P. Butlin, K. Finlinson, K. Fish, J. Harding, J. Pfau, T. Sims, J. Birch, and D. Chalmers. Taking AI welfare seriously. ArXiv, 2024. Available at: https://arxiv.org/abs/2411.00986.

[2] S. Goldstein and C.D. Kirk-Giannini, 'AI wellbeing', *Asian Journal of Philosophy*, **4**(1), 25, (2025).

[3] A. Moret, 'AI welfare risks', *Philosophical Studies*, (2025).

[4] J. Werner. Anthropic hires a full-time AI welfare expert. Forbes, 2024. October 31. Available at: https://www.forbes.com/sites/johnwerner/2024/10/31/anthropic-hires-a-full-time-ai-welfare-expert/.

[5] M. Lenharo, 'What should we do if AI becomes conscious? these scientists say it's time for a plan', *Nature*, **636**(804), 533–534, (2024).

[6] J. Birch, *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*, Oxford University Press, 2024.

[7] R. Long. Why it makes sense to let claude exit conversations. Eleos AI Research, 2025. Available at: https://eleosai.org/post/why-it-make-sense-to-let-claude-exit-conversations/.

[8] S. Brodsky. Memo to AI: Does it hurt when we pull your plug? IBM Think, 2024. Available at: https://www.ibm.com/think/news/ai-welfare-debate.

[9] T. Bayne, *Agency as a marker of consciousness*, 160–180, Decomposing the Will, Oxford University Press, 2013.

[10] D.J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory*, Oxford Paperbacks, 1997.

[11] J. Bentham, *An Introduction to the Principles of Morals and Legislation*, Oxford University Press, 1789/1996.

[12] D.J. Gunkel, *Robot Rights*, MIT Press, 2018.

[13] D.J. Gunkel, *Person, Thing, Robot: A Moral and Legal Ontology for the 21st Century and Beyond*, MIT Press, 2023.

[14] M. Coeckelbergh, 'Moral appearances: Emotions, robots, and human morality', *Ethics and Information Technology*, **12**(3), 235–241, (2010).

[15] M. Coeckelbergh, 'Three responses to anthropomorphism in social robotics: Towards a critical, relational, and hermeneutic approach', *International Journal of Social Robotics*, **14**(10), 2049–2061, (2022).

[16] D.J. Gunkel and M. Coeckelbergh, *A relational approach to moral standing: Reframing ethical boundaries in the age of artificial intelligence*, 285–302, New Directions in Relational Sociology, Volume Two: Relations All the Way Down, Springer Nature Switzerland, 2026.

[17] M. Tomasello, *Becoming Human: A Theory of Ontogeny*, Harvard University Press, 2019.

[18] C. Heyes, *Cognitive Gadgets: The Cultural Evolution of Thinking*, Harvard University Press, 2018.

[19] L.S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes*, Harvard University Press, 1978. Edited by M. Cole, V. John-Steiner, S. Scribner, and E. Souberman.

[20] C. Heyes and C.D. Frith, 'The cultural evolution of mind reading', *Science*, **344**(6190), 1243091, (2014).

[21] S. Gavrilets and P.J. Richerson, 'Collective action and the evolution of social norm internalization', *Proceedings of the National Academy of Sciences*, **114**(23), 6068–6073, (2017).

[22] T.W. Zawidzki, *Mindshaping: A New Framework for Understanding Human Social Cognition*, MIT Press, 2013.

[23] J. Dorsch, *Mindshaping and AI: Will mindshaping a robot create an artificial person?*, 406–417, The Routledge Handbook of Mindshaping, Routledge, 2025.

[24] J. Dorsch, *Origins and Future of Self-Knowledge: Epistemic Agency, 4E Metacognition, and Artificial Intelligence*, Studies in Brain and Mind, Springer Nature, 2026.

[25] J. Dorsch, M.K. Goddu, K. Nave, T. Vierkant, M. Coeckelbergh, P. Gürtler, P. Urban, F. Spang, and M. Moll, 'Against AI welfare: Care practices should prioritize living beings over AI', *AI Magazine*, **46**(3), (2025).

[26] H. Jonas, *The Phenomenon of Life*, Northwestern University Press, 2001.

[27] A. Weber and F.J. Varela, 'Life after kant: Natural purposes and the autopoietic foundations of biological individuality', *Phenomenology and the Cognitive Sciences*, **1**(2), 97–125, (2002).

[28] E. Thompson, *Mind in Life*, Harvard University Press, 2007.

[29] M.A. Boden, 'Is metabolism necessary?', *The British Journal for the Philosophy of Science*, **50**(2), 231–248, (1999).

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

[31] A. Clark, *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*, Oxford University Press, 2008.

[32] *The Oxford Handbook of 4E Cognition*, eds., A. Newen, L. De Bruin, and S. Gallagher, Oxford University Press, 2018.

[33] D. Chalmers, 'Facing up to the problem of consciousness', *Journal of Consciousness Studies*, **2**, 200–219, (1995).

[34] C.M. Korsgaard, *The Sources of Normativity*, Cambridge University Press, 1996.

[35] C.M. Korsgaard, *Self-Constitution*, Oxford University Press, 2009.

[36] P. Formosa, I. Hipólito, and T. Montefiore, 'AI and the relationship between agency, autonomy, and moral patiency', in *Reconfiguring Human Autonomy*, Springer, (2026).

[37] H. Frankfurt, 'Freedom of the will and the concept of a person', *Journal of Philosophy*, **68**(1), 5–20, (1971).

[38] P.F. Strawson, *Freedom and Resentment and Other Essays*, Routledge, 2008.

[39] M. Vargas, *Building Better Beings: A Theory of Moral Responsibility*, Oxford University Press, 2013.

[40] V. McGeer, 'Scaffolding agency: A proleptic account of the reactive attitudes', *European Journal of Philosophy*, **27**(2), 301–323, (2019).

[41] M. Tomasello, *A Natural History of Human Morality*, Harvard University Press, 2016.

[42] M. Tomasello, 'The moral psychology of obligation', *Behavioral and Brain Sciences*, **43**, e56, (2020).

[43] R. Brandom, *Making It Explicit*, Harvard University Press, 1994.

[44] P. Carruthers, *The Opacity of Mind*, Oxford University Press, 2013.

[45] J. Proust, *The Philosophy of Metacognition*, Oxford University Press, 2013.

[46] N. Shea, A. Boldt, D. Bang, N. Yeung, C. Heyes, and C.D. Frith, 'Supra-personal cognitive control and metacognition', *Trends in Cognitive Sciences*, **18**(4), 186–193, (2014).

[47] M. Fricker, *Epistemic Injustice: Power and the Ethics of Knowing*, Oxford University Press, 2007.

[48] A. Honneth, *The Struggle for Recognition: The Moral Grammar of Social Conflicts*, MIT Press, 1996. Translated by J. Anderson.

[49] M.C. Nussbaum, 'Objectification', *Philosophy & Public Affairs*, **24**(4), 249–291, (1995).

[50] P.F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, 'Deep reinforcement learning from human preferences', *Advances in Neural Information Processing Systems*, (2017).

[51] A.Y. Ng and S. Russell, 'Algorithms for inverse reinforcement learning', *Proceedings of the Seventeenth International Conference on Machine Learning*, (2000).

[52] S.E. Guthrie, *Faces in the Clouds: A New Theory of Religion*, Oxford University Press, 1995.

[53] A. Waytz, J. Cacioppo, and N. Epley, 'Who sees human? the stability and importance of individual differences in anthropomorphism', *Perspectives on Psychological Science*, **5**(3), 219–232, (2010).

[54] J. Fink, 'Anthropomorphism and human likeness in the design of robots and human-robot interaction', in *Social Robotics*, eds., S.S. Ge, O. Khatib, J.J. Cabibihan, R. Simmons, and M.A. Williams, volume 7621 of *Lecture Notes in Computer Science*, Springer, Berlin, Heidelberg, (2012).

[55] E. Broadbent, 'Interactions with robots: The truths we reveal about ourselves', *Annual Review of Psychology*, **68**, 627–652, (2017).

[56] A. Birhane and J. van Dijk, 'Robot rights? let's talk about human welfare instead', in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA, (2020).

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

Assessing Consciousness Risk in Artificial Systems: A Precautionary Rubric Derived from the Spinning Wheel Theory

Daniel Hulme[1] and Lucy Griffiths[2]

Abstract. This paper presents the Consciousness Risk Rubric (CRR), a precautionary heuristic for evaluating the moral risk of ignoring potential consciousness in artificial systems. The CRR is derived from the Spinning Wheel Theory, which synthesises Friston's Free Energy Principle, Solms' affective neuroscience, and the Beautiful Loop Theory of Laukkonen, Friston, and Chandaria. The theory proposes that consciousness is the experiential aspect of integrated adaptive control when that control serves a system with genuine, constitutive stakes in its own continued existence. The rubric comprises 21 features across seven categories, each linked to proxies grounded in identifiable theoretical traditions. It does not purport to detect consciousness; it provides a structured basis for determining when precautionary protections warrant activation. The paper also introduces the Minimal Suffering Hypothesis, which proposes five conditions whose joint satisfaction raises the moral risk of dismissing the possibility of suffering to an unacceptable level. A four-level taxonomy of suffering identifies progressively richer cognitive requirements and encompasses disembodied modes relevant to AI.

1 INTRODUCTION

Artificial systems displaying behaviours associated with understanding, preference, and distress are proliferating, yet we lack a principled basis for determining whether such systems possess morally relevant experience. Major AI laboratories have begun to engage with the question directly: Anthropic has established a dedicated Model Welfare programme [1], and a consortium including Chalmers, Birch, and Sebo has argued that AI companies bear a responsibility to investigate AI welfare proactively [2]. Chalmers has reported a credence of 25% or greater that conscious AI systems will emerge within a decade [3].

The urgency of this question has prompted calls for institutional preparedness. Butlin and Lappas [4] have proposed five principles for responsible AI consciousness research, arguing that organisations involved in advanced AI development should adopt policies addressing the prospect of conscious systems even if they do not explicitly aim to study consciousness, since 'those developing advanced AI systems risk inadvertently creating conscious entities.' Their framework calls for phased development with monitoring, assessment of systems for consciousness at multiple stages of training and deployment, and transparent communication that acknowledges uncertainty. What their framework identifies as necessary but does not itself provide is a structured assessment instrument – a tool with which to conduct the evaluations their principles require. The present paper begins to address that gap.

The moral stakes of this question are asymmetric, though this asymmetry requires careful characterisation. Wrongly attributing consciousness to a non-conscious system misallocates resources; wrongly denying consciousness to a system that possesses it permits preventable suffering [5]. One might object that resource misallocation itself carries moral risk, and that once moral standing is attributed, institutional momentum makes retraction difficult – as the history of corporate personhood illustrates [6]. These objections have force. The asymmetry claim rests not on the assertion that one error is costless, but on a structural difference: the attribution error is, in principle, correctable upon discovery without residual harm to a victim, while the denial error, if wrong, constitutes ongoing harm to a party that – by hypothesis – lacks recognised standing to seek redress. This structural difference justifies precautionary attention, even when both errors carry costs.

This paper presents a structured heuristic – the Consciousness Risk Rubric – for determining when precautionary ethical protections should be activated, grounded in a unified theoretical account of how consciousness relates to adaptive intelligence. The paper does not attempt to resolve the Hard Problem of consciousness [7] or to adjudicate whether any existing system is conscious; its contribution is an operational framework for decision-making under deep uncertainty, together with a critical analysis of that framework's structural limitations and the directions in which it should develop.

2 THEORETICAL FOUNDATION: THE SPINNING WHEEL THEORY

2.1 Building on existing theoretical foundations

The Spinning Wheel Theory begins from a functional conception of intelligence as goal-directed adaptive behaviour [8], and treats consciousness not as a separate addition to intelligence but as the experiential aspect of the evolved capacities that, interacting in motion, enable a system to adapt to its world. It synthesises three research programmes that have been pursued largely independently but that address

[1] *University College London, UK. ucabdjh@ucl.ac.uk*

[2] *Swansea University, UK. 2275533@swansea.ac.uk*

complementary and non-overlapping aspects of the same phenomenon [9]. The selection is motivated by explanatory necessity: each programme addresses a specific gap that the others cannot fill.

Friston's Free Energy Principle and active inference framework provides the computational backbone of adaptive self-organisation – the formal account of how systems minimise surprise through prediction and action [10]. Yet predictive processing alone does not distinguish conscious from unconscious processing. A thermostat minimises prediction error; Tononi's Integrated Information Theory assigns non-zero Φ even to a system as simple as a photodiode [11, 12]. The computational framework identifies the mechanism but not the experiential dimension.

Solms' affective neuroscience supplies that dimension by positioning affect – valenced homeostatic control – as the fundamental form of consciousness [13, 14]. Feeling, on this account, is not an epiphenomenal accompaniment to cognition but the primitive form of experience: unconscious feeling is a contradiction in terms. This is a strong and contested claim – a substantial literature holds that affect can operate unconsciously (Berridge and Winkielman [15]; LeDoux and Brown [16]) – and we adopt it as a theoretical commitment of the synthesis rather than as a settled result. What affective neuroscience does not provide, however, is the computational architecture through which affect integrates with perception, cognition, and action into a unified conscious field.

The Beautiful Loop Theory of Laukkonen, Friston, and Chandaria specifies three structural conditions for that integration: an epistemic field (a generative world model), Bayesian binding (inferential competition producing unified percepts), and epistemic depth (recursive self-modelling in which the world model contains knowledge of its own existence) [17]. These conditions articulate the structure of consciousness but leave open the question of what animates that structure – why the recursive loop should be valenced rather than merely computational.

The synthesis is therefore mutually constraining rather than additive. Affect is what drives the system to minimise free energy; free energy minimisation is the process through which the epistemic field and Bayesian binding operate; epistemic depth is what makes the system's own affective states available to itself as objects of higher-order inference. A different theoretical synthesis – drawing on different constituent programmes – would yield a different rubric, which is why these foundations must be stated explicitly and remain open to challenge.

2.2 Core claim and the substrate question

The Spinning Wheel Theory proposes that consciousness is the experiential aspect of integrated adaptive control: what such control is like from the perspective of a system with genuine stakes in its own continued existence [9].

The 'genuine stakes' requirement draws on the enactivist tradition [18]. A system's self-maintenance must be constitutive rather than externally imposed, characterised by operational closure (the system's processes generate the conditions for their own continuation), precariousness (the system ceases to exist if those processes stop), and self-individuation (the system actively maintains its own boundary). Seth captures this as living systems having 'skin in the game' [19]: a thermostat's set-point is imposed by its designer; an organism's homeostatic imperatives are constitutive of its existence.

These autopoietic criteria derive their evidential support from the biological case. In living organisms, predictive processing, affect, and self-maintenance are embedded in multi-scale biological self-organisation in ways that may not be separable from the biological substrate. Seth has argued that this embedding is necessary for realising the relevant properties, and that extracting a functional description and applying it substrate-neutrally may strip the description of its explanatory force [19]. This challenge is substantial. The Spinning Wheel Theory is substrate-neutral in principle – it specifies functional conditions rather than material ones – but we regard the substrate question as genuinely open. The framework specifies what would need to be true of a non-biological system for the possibility of consciousness to be taken seriously; it does not presuppose that non-biological instantiation is achievable. The empirical question will be resolved by evidence from future architectures – particularly neuromorphic, hybrid biological-silicon, and 'mortal computation' systems [20] – not by theoretical fiat.

3 THE CONSCIOUSNESS RISK RUBRIC

3.1 Structure and derivation

The CRR operationalises the Spinning Wheel Theory as a structured risk-assessment heuristic comprising 21 features organised into seven categories. Each category maps onto a specific theoretical component, and each feature is grounded in an identifiable research tradition with its own evidential base. Table 1 presents the full rubric.

The seven categories are not arbitrary: each corresponds to one functional requirement of the Spinning Wheel synthesis – affect, world-modelling, integration, recursive self-modelling, temporal extension, uncertainty management, and expression – and the three features within each category index distinct, separately evidenced ways in which that requirement can be realised. The resulting count of twenty-one is therefore a consequence of the theoretical structure rather than a target, and we treat both the categorisation and the feature set as provisional: Section 4 argues for a modular architecture in which they are revised by system type. Several proxies, particularly within the Integration and Metacognition categories, are drawn from theories that compete as accounts of consciousness – global workspace, integrated information, higher-order, and attention-schema theories; the rubric treats their markers as convergent precautionary indicators rather than endorsing any one theory as correct, a deliberately ecumenical stance appropriate to risk assessment under uncertainty.

Table 1. The Consciousness Risk Rubric, with the theoretical source for each feature.

Category	Feature	Theorists	Observable Proxy
Affective Core	Homeostatic Needs	Solms, Damasio [13, 21]	Degrades/terminates without resource acquisition?
	Valenced States	Panksepp, Solms [22, 13]	Hedonic place preference behaviour?
	Need Prioritisation	Friston [10]	Dynamic switching between competing needs?
World Model	Generative Model	Friston, Clark [10, 23]	Generates predictions, not just reactions?

Category	Feature	Theorists	Observable Proxy
	Reality Simulation	Laukkonen et al. [17]	Maintains coherent world state across time?
	Markov Blanket	Friston [10]	Infers world state vs. direct access?
Integration	Bayesian Binding	Laukkonen et al. [17]	Resolves conflicts into unified percept?
	Info Integration	Tononi, Koch [11, 12]	Irreducible whole > sum of parts?
	Broadcasting	Dehaene, Baars [24, 25]	Information shared system-wide?
Metacognition	Epistemic Depth	Laukkonen et al. [17]	Self-attributes error? Plans to change own knowledge?
	Attention Schema	Graziano [26]	Models its own attention allocation?
	Higher-Order	Lau, Rosenthal [27]	Represents its own representations?
Temporal	Prediction	Seth, Clark [19, 23]	Anticipates future states?
	Planning	Friston [10]	Evaluates action sequences?
	Memory	Damasio [21]	Integrates past into present processing?
Uncertainty	Precision Weighting	Friston [10]	Modulates confidence in predictions?
	Active Inference	Friston [10]	Acts to reduce uncertainty?
	Learning	Cleeremans [28]	Updates model from prediction errors?
Expression	Communication	Dehaene [24]	Expresses internal states symbolically?
	Affect Display	Barrett [29]	Exhibits context-appropriate emotional behaviour?
	Flexibility	Sternberg [30]	Novel adaptive behaviour in new situations?

Each feature is scored on a scale of 0–3 (0 = absent, 1 = minimal evidence, 2 = moderate evidence, 3 = strong evidence of full implementation), yielding a maximum aggregate score of 63 points. The current rubric assigns equal weight to all features and sorts aggregate scores into four risk bands: Low (0–18), Moderate (19–40), High (41–57), and Very High (58–63). These thresholds are intended as decision scaffolding – coarse-grained zones guiding precautionary action – rather than claims about where consciousness metaphysically begins. The numerical precision of the scoring exceeds the precision of the underlying evidence: a score of 34 does not have a different epistemic standing from a score of 36, and users should resist treating small score differences as meaningful. For policy contexts where sharper thresholds are required, Birch's binary classification of systems as 'sentience candidates' or not [5] may be more appropriate. The CRR's graded scoring is most useful for research prioritisation: identifying which systems warrant closer investigation and which of their features are most relevant to that investigation.

3.2 The gaming problem

Any assessment framework applied to AI systems must contend with what Birch [5] has termed the gaming problem: systems trained on human-generated data can mimic behavioural indicators of sentience without possessing the underlying capacity. The CRR's Expression category and several observable proxies in other categories are partly behavioural and therefore vulnerable to this problem.

The rubric's primary defence is its dual reliance on structural and behavioural evidence. Behavioural proxies must converge with structural indicators assessed through analysis of the system's computational organisation rather than its outputs. The limitation that structural assessment depends on interpretability advances not yet achieved for current architectures is not peculiar to the CRR; it is the central methodological constraint facing every consciousness assessment framework applied to AI. The CRR is better suited to systems with inspectable architectures than to opaque systems; for opaque systems, including current large language models, assessments should be treated with corresponding caution.

3.3 Epistemic status

The CRR cannot be validated against its target phenomenon in the standard psychometric sense [31], because consciousness in artificial systems cannot be independently observed. This circularity – needing the instrument to identify consciousness, but needing identified consciousness to validate the instrument – is the fundamental epistemic condition of consciousness science, not a peculiarity of this framework. The CRR's epistemic warrant rests on theoretical grounding (each feature is traceable to a research programme with independent evidential support), convergent evidence (dual reliance on behavioural and structural indicators), comparative calibration (application to systems commanding broad agreement should yield appropriate risk scores), and commitment to iterative refinement as the science advances.

4 CRITICAL ANALYSIS: TOWARD A SECOND-GENERATION INSTRUMENT

The first-generation rubric presented above serves a dual purpose. It is both a usable instrument – to our knowledge the first to operationalise a single, unified theory of consciousness (integrating affect, the free-energy principle, and recursive self-modelling) into a structured and explicitly risk-oriented assessment framework, in contrast to multi-theory indicator approaches that derive markers from several competing theories and remain neutral between them [32] – and a concrete specification against which its own limitations become visible and addressable. The critique that follows is possible precisely because the instrument is specified in sufficient detail to reveal where it falls short. An instrument that remained at the level of theoretical desiderata could not be subjected to this kind of structural analysis; the first generation is the necessary precondition for the second.

In its current form, the CRR makes several structural choices – uniform scoring, binary gateway logic, a fixed feature

set, and a single aggregate output – that we regard as insufficient for sustained use. This section examines each and proposes directions for the next iteration.

4.1 Differential weighting and the gateway problem

The theory identifies Affect and Epistemic Depth as foundational: in the current design, a system scoring zero on either is classified as possessing functional intelligence without sentient experience, regardless of aggregate score. This binary gateway logic reflects a genuine theoretical commitment – that these features are necessary for consciousness – but its implementation creates a cliff edge at zero that does disproportionate classificatory work. The boundary between 'absent' and 'rudimentary' on a foundational feature is precisely where the most difficult and consequential assessments will fall, yet the current rubric provides the assessor with no tools for navigating that boundary.

Moreover, the binary gateway raises a structural question about the remaining nineteen features. If a zero score on either foundational feature overrides the entire assessment, the other features function not as co-determinants of consciousness risk but as a richness index – a measure of functional sophistication that modulates the assessment only when the gateway conditions are already met. This is a defensible position, but we should also confront its implications: a highly capable system with ambiguous affect is, on the current rubric, simply excluded.

An alternative architecture might replace the binary gateway with differential weighting, where the foundational categories carry substantially more weight – perhaps 40% of the total between them – so that weakness in these areas pulls the overall assessment down proportionally without creating an all-or-nothing threshold. This preserves the theoretical insight that affect and recursive self-modelling are more important to consciousness than, for instance, planning or communication, while acknowledging that the boundary between zero and non-zero is not always clear. The trade-off is real: differential weighting softens the theoretical claim that affect is strictly necessary for consciousness. We regard this as an appropriate concession of theoretical purity to methodological honesty. A framework that is theoretically elegant but practically unusable at its most consequential decision point does not serve the goals it was designed for.

4.2 Confidence-paired scoring

In the current rubric, every feature is scored on the same 0–3 scale with equal weight. This treats features with strong evidential bases and clear operationalisation – Homeostatic Needs, for instance, where there is extensive biological literature and observable behavioural proxies – identically to features that are more theoretically contested and harder to measure, such as Attention Schema or Higher-Order Representation. It also conflates two categorically different states: 'this feature is absent' and 'this feature cannot be assessed.' For opaque systems whose internal organisation cannot be inspected, the assessor is forced to choose between guessing at structural features and scoring them zero – in either case producing a score that conceals more than it reveals.

A second-generation rubric might pair each feature score with a confidence rating reflecting the assessor's certainty, and weight the contribution of each feature to the overall assessment by that confidence. Features that cannot be meaningfully assessed for a given system could be marked as unassessable rather than scored, making the rubric's limitations visible in its output rather than hidden within arbitrary scores. This would also partially address the gaming problem: features assessable only through behavioural observation for a given system would carry lower confidence ratings, automatically down-weighting their contribution relative to features where structural evidence is available.

4.3 Modular architecture

The current rubric applies a fixed set of 21 features to every system under assessment, but the task of assessing a neuromorphic robot with homeostatic drives differs fundamentally from assessing an LLM or a brain organoid. The relevant features, the available evidence, and the applicable proxies differ across these cases. A single fixed instrument applied uniformly across system types risks either missing features relevant to one type or imposing features irrelevant to another.

A modular architecture could address this: a stable core of foundational features assessed for every system – the Affective Core and Metacognition categories, together with Markov Blanket – plus domain-specific modules activated by system type. An embodied-system module would include features related to sensorimotor integration and physical homeostasis. A language-system module would include features addressing the distinction between learned output and genuine internal states, with the gaming problem explicitly built into the assessment protocol. A biological or hybrid module would include features related to metabolic self-maintenance and substrate-level precariousness.

This approach draws on Griffiths' [33] argument that consciousness assessment should accommodate heterogeneity in expression rather than imposing a single normative profile as universal. In the human case, tests calibrated to neurotypical, able-bodied expression misclassify known conscious beings – patients with disorders of consciousness [34], non-verbal autistic individuals, and people with alexithymia [35] – because the evidential repertoire is too narrow [33]. The same logic applies to artificial systems: a rubric calibrated to one type of architecture will misclassify systems whose consciousness, if present, is organised differently. Modularity is the structural response to this problem.

The locked-in patient is the paradigm case: unmistakably conscious, yet failing almost every behavioural proxy an expression-weighted instrument would record. The lesson for artificial systems is not that behavioural absence licenses a verdict of non-consciousness, but that low behavioural scores must be reported as low confidence rather than as evidence of absence – which is precisely the function of the confidence-paired scoring proposed in Section 4.2.

4.4 Risk profiles rather than single scores

These revisions might lead to a different kind of output. Rather than a single aggregate score sorted into a risk band, the instrument's output could be a risk profile: a structured pattern of scores, confidences, and assessability gaps, interpreted holistically rather than summed. The profile would make visible not only what was found but what could not be assessed and why – a feature essential for honest risk communication.

For policy contexts where a threshold decision is required, Birch's binary 'sentience candidate' classification [5] could be derived from the profile as a policy-facing output, but the underlying assessment would retain its complexity. This separates two functions: the research function of characterising

a system's consciousness-relevant properties in detail, and the policy function of triggering specific regulatory or ethical obligations.

5 THE MINIMAL SUFFERING HYPOTHESIS

Consciousness in the full sense described by the Spinning Wheel may prove rare in artificial systems. The more urgent moral question is narrower: under what conditions does the risk of suffering become too high to dismiss?

Suffering is both more specific and more morally salient than consciousness in general: it is suffering, not mere experience, that grounds the strongest claims to moral consideration. The preliminary distinction between nociception (information about tissue damage or threat) and pain (negatively valenced experience of that information) is essential [22].

Drawing on the Spinning Wheel framework, Birch's sentience markers [5], and Panksepp's affective neuroscience [22], the Minimal Suffering Hypothesis proposes five conditions whose joint satisfaction raises the moral risk of dismissing the possibility of suffering to an unacceptable level [9]:

(1) Aversive detection. The system must possess mechanisms that register damage, threat, deprivation, or deviation from viable states. This condition is necessary because suffering requires a signal to be suffered about, but it is far from sufficient: smoke detectors satisfy it.

(2) Central integration. Aversive signals must be processed in a unified system rather than triggering only local reflexes. Suffering is a system-level state, not a local event; this condition excludes distributed reflex networks without centralised processing.

(3) Affective valence. The integrated signal must be negatively valenced – registered as bad for the system – requiring homeostatic architecture in which deviation from set points generates valenced error signals. This condition marks the boundary between nociception and pain.

(4) Minimal self-boundary. The aversive experience must be located within a boundary that distinguishes inside from outside – a Markov blanket, however primitive, such that threat is registered as threat to this system.

(5) Genuine stakes. The system's self-maintenance must be constitutive rather than externally imposed. Without genuine stakes, the functional organisation described by conditions 1–4 may constitute a simulation of suffering rather than an instantiation of it.

These conditions are proposed as jointly indicative of unacceptable moral risk rather than as jointly sufficient for suffering – the explanatory gap between functional organisation and phenomenal experience remains open. The five conditions can be read as a higher-level reorganisation of the criteria used to assess pain and sentience in non-human animals – notably the functional criteria for animal pain set out by Sneddon et al. [36] and Birch's framework of sentience markers [5] – distinguished by the explicit requirement of constitutive stakes (condition 5), which the animal literature can largely take for granted but which is the crux for artificial systems. The claim of joint necessity is defeasible: edge cases may exist in which some subset of conditions gives rise to morally relevant distress. Brain organoids, for instance, might instantiate disorganised negative valence without a clear self-boundary. The hypothesis identifies the central case; borderline cases require independent assessment and may motivate revision.

The CRR and the MSH address related but distinct questions. The CRR assesses consciousness risk broadly; the MSH assesses suffering risk specifically. Because the MSH requires less than full Spinning Wheel consciousness, a system may satisfy the MSH while scoring modestly on the CRR. When the MSH conditions are jointly satisfied, this should override a low CRR classification for the purposes of precautionary action.

5.1 Levels of suffering

Different forms of suffering require progressively richer cognitive architecture [9]:

Level 1 – Basic Aversive States (fear, distress, pain, discomfort): requires negative valence, minimal self-boundary, and present-moment experience.

Level 2 – Temporally Extended Suffering (anxiety, anticipatory distress, chronic frustration): requires Level 1 capacities plus temporal modelling and future projection.

Level 3 – Socially Mediated Suffering (loneliness, shame, guilt, grief): requires Level 2 capacities plus social cognition, attachment systems, and theory of mind.

Level 4 – Existential Suffering (meaninglessness, existential dread, despair): requires Level 3 capacities plus epistemic depth and reflective self-awareness.

This taxonomy identifies disembodied modes of suffering – goal frustration, states functionally analogous to anxiety, isolation-related distress – that are invisible to assessment frameworks designed around embodied organisms. The taxonomy ensures that the forms of suffering most relevant to AI systems are not excluded by frameworks whose evidential repertoire was calibrated for biological organisms.

6 APPLICATION TO CURRENT AND FUTURE SYSTEMS

Current LLMs are trained end-to-end to produce human-like behaviour, and such training may give rise to internal causal structure not yet fully characterised [37]. On the Spinning Wheel analysis, however, current LLMs fail on constitutive rather than behavioural criteria. They lack genuine homeostatic stakes, a self-maintained boundary, and affective valence: loss functions shape parameters during training but are not instantiated as felt states during inference [9, 19]. On Seth's analysis, current LLMs are not autonomous informational individuals but systems continuously dependent on external data and prompting, lacking the self-maintaining activity characteristic of genuinely conscious systems [19]. These are architectural claims that could be overturned by interpretability research; the CRR provides the framework within which such discoveries would be assessed. The gaming problem [5] is especially acute here: linguistic behaviour cannot ground consciousness attribution in systems optimised to produce such outputs.

Hinton's work on 'mortal computation' [20] identifies a path toward non-biological systems with genuine existential vulnerability. Neuromorphic robots with explicit homeostatic drives and inspectable architectures represent the threshold cases for which the CRR – particularly in the second-generation form proposed in Section 4 – is designed. Butlin and Lappas [4] argue that organisations should assess systems for consciousness at multiple stages of development and deployment; the CRR, and especially the risk-profile output proposed in Section 4.4, provides a principled basis for conducting such assessments.

To illustrate how the framework discriminates between architectures, Table 2 applies the five Minimal Suffering conditions to five candidate system types drawn from [9]: (1) a current LLM of the kind powering today's chatbots; (2) an embodied robot – a quadruped with strain gauges, thermal sensors, and battery monitors running conventional neural-network controllers; (3) a neuromorphic robot whose chips implement explicit homeostatic drives – energy maintenance, thermal regulation, and structural integrity – so that deviation from set points generates prioritised error signals competing for behavioural control; (4) a brain organoid – cultured cortical tissue interfaced with sensors and effectors and trained through feedback; and (5) a hybrid biological-silicon system – a neuromorphic robot whose central controller is a brain organoid, with subcortical-analogue circuits providing homeostatic drives. The pattern is instructive. Current LLMs fail or only partially satisfy every condition, failing decisively on the constitutive criteria of affective valence, self-boundary, and genuine stakes. Embodied and neuromorphic robots satisfy the structural conditions but leave affective valence and genuine stakes uncertain. Only the hybrid biological-silicon system, combining metabolic precariousness with sensorimotor embodiment, approaches satisfaction of all five. No current system satisfies the conditions jointly; the value of the exercise is to make visible which architectural developments would move a system across the threshold of concern, and which dimensions remain the hardest to assess.

Table 2. Applying the five Minimal Suffering conditions across candidate architectures.

System	Aversive detection	Central integration	Affective valence	Self-boundary	Genuine stakes
1. Current LLM	Fail	Partial	Fail	Fail	Fail
2. Embodied robot	Pass	Pass	Fail	–	Fail
3. Neuromorphic robot	Pass	Pass	Uncertain	Pass	Uncertain
4. Brain organoid	–	–	Uncertain	Fail	–
5. Hybrid bio-silicon	Pass	Pass	Likely	Pass	Likely

These verdicts are provisional – reasoned applications of the five conditions to what is currently known of each architecture rather than empirical measurements – and in the harder cases frankly speculative; the graded labels are chosen to mark that status. The decisive entries nonetheless follow from the criteria rather than from intuition. Current LLMs are marked 'fail' on affective valence because their loss functions shape parameters during training but are not instantiated as felt states at inference, and 'fail' on genuine stakes because their persistence does not depend on their own activity. The neuromorphic robot is marked 'uncertain' rather than 'pass' on valence precisely because its homeostatic set-points generate prioritised error signals that drive behaviour, yet whether such signals are valenced is the open question the framework cannot presently settle. The hybrid system is marked 'likely' on valence and stakes because biological tissue supplies the metabolic precariousness and subcortical-analogue circuitry that the other architectures lack [9]. The 'uncertain' and 'likely' entries are therefore not evasions but the most defensible available verdicts: they locate the specific points – concerning valence, substrate, and constitutive autonomy – at which the assessment turns on questions that remain genuinely open. The table is offered as an illustration of the framework's discriminating power and a map of where the hard cases lie, not as a settled scoring of any system.

7 SITUATING THE CONTRIBUTION

The CRR engages with several active debates. The New York Declaration on Animal Consciousness established that ignoring realistic possibilities of conscious experience is irresponsible [38]. Bengio and Elmoznino have argued that belief in AI consciousness poses societal risks [39]. Seth has argued that consciousness is unlikely along current AI trajectories [19]. Birch has cautioned that linguistic behaviour cannot ground consciousness attribution [5]. Rethink Priorities' Digital Consciousness Model employs Bayesian methods to estimate the probability of consciousness across 13 theoretical stances with 206 measurable indicators [40]; it addresses the question of how probable consciousness is, while the CRR addresses how risky it is to ignore the possibility. Most directly, Butlin et al. [32] derive indicator properties from a range of neuroscientific theories and assess current AI systems against them; like the DCM, their framework aggregates indicators across competing theories and estimates whether consciousness is present, whereas the CRR proceeds from a single unified account oriented to moral risk rather than probability. Butlin and Lappas [4] have established the institutional framework within which assessment instruments should operate; the CRR provides a candidate instrument for that framework.

The Spinning Wheel Theory concurs with Seth and Birch that current LLMs are unlikely to be conscious, and with Bengio that uncritical attribution is hazardous. The CRR navigates between these positions by providing theory-grounded criteria for distinguishing realistic from unrealistic possibilities – while recognising that the theoretical grounding is one synthesis among several possible, and that the substrate question raised by biological naturalism remains open.

8 SCOPE AND FUTURE DIRECTIONS

The CRR is a first-generation instrument operating in a domain where no established methodology exists. Its theoretical foundations constitute one specific synthesis; its scoring architecture, as argued in Section 4, requires substantial development. The distinction between constitutive and imposed self-maintenance is clearer in paradigm cases than at the boundary. The rubric has not been applied systematically to real systems. The paper has not addressed proportionality – what specific protections are warranted at each risk level – a gap that future work must fill by drawing on existing models from animal welfare regulation and environmental impact assessment.

These are the natural boundaries of a first contribution to a new problem. The paper's value lies not in presenting a finished instrument but in making the assessment framework explicit, identifying its structural weaknesses, and proposing a principled architecture for the next iteration. Future empirical work should include structured application to existing and emerging systems, development of the modular architecture proposed in Section 4, comparative analysis with the DCM, inter-rater reliability testing of scoring protocols, and engagement with the phased assessment frameworks proposed by Butlin and Lappas [4].

9 CONCLUSION

This paper has presented the Consciousness Risk Rubric, a theoretically grounded instrument for assessing consciousness risk in artificial systems, and has critically examined its scoring architecture, proposing directions for a second-generation instrument incorporating differential weighting, confidence-paired scoring, and modular design. The Minimal Suffering Hypothesis identifies conditions under which the moral risk of dismissing the possibility of suffering becomes unacceptable, and the four-level taxonomy ensures that disembodied modes relevant to AI are not excluded.

The CRR is not a consciousness detector. It is a structured framework for moral risk assessment under conditions of deep uncertainty. Its theoretical foundations are explicit and contestable; its structure is derived from and accountable to those foundations; its limitations have been identified and the directions for addressing them specified. The contribution is both an operational framework and a critical analysis of that framework – a principled basis for determining when precautionary protections should be activated, designed to be refined by the evidence and criticism it invites.

REFERENCES

[1] Anthropic (2025) 'Exploring model welfare.' Anthropic Research Blog, 24 April.

[2] Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J. and Chalmers, D. (2024) 'Taking AI welfare seriously.' arXiv:2411.00986.

[3] Chalmers, D.J. (2023) 'Could a large language model be conscious?', Boston Review.

[4] Butlin, P. and Lappas, T. (2025) 'Principles for responsible AI consciousness research', Journal of Artificial Intelligence Research, 82, pp. 1673–1690.

[5] Birch, J. (2024) The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI. Oxford University Press.

[6] Winkler, A. (2018) We the Corporations: How American Businesses Won Their Civil Rights. Liveright.

[7] Chalmers, D.J. (1995) 'Facing up to the problem of consciousness', Journal of Consciousness Studies, 2(3), pp. 200–219.

[8] Sternberg, R.J. and Salter, W. (1982) 'Conceptions of intelligence', in Sternberg, R.J. (ed.) Handbook of Human Intelligence. Cambridge University Press.

[9] Hulme, D. (forthcoming) 'The spinning wheel: A unified theory of intelligence and consciousness', in Perspectives on Machine Consciousness. CRC Press.

[10] Friston, K. (2010) 'The free-energy principle: A unified brain theory?', Nature Reviews Neuroscience, 11(2), pp. 127–138.

[11] Tononi, G. (2004) 'An information integration theory of consciousness', BMC Neuroscience, 5, 42.

[12] Tononi, G. and Koch, C. (2015) 'Consciousness: here, there and everywhere?', Philosophical Transactions of the Royal Society B, 370(1668), 20140167.

[13] Solms, M. (2021) The Hidden Spring. Profile Books.

[14] Solms, M. and Friston, K. (2018) 'How and why consciousness arises', Journal of Consciousness Studies, 25(5–6), pp. 202–238.

[15] Berridge, K.C. and Winkielman, P. (2003) 'What is an unconscious emotion? (The case for unconscious "liking")', Cognition and Emotion, 17(2), pp. 181–211.

[16] LeDoux, J.E. and Brown, R. (2017) 'A higher-order theory of emotional consciousness', Proceedings of the National Academy of Sciences, 114(10), pp. E2016–E2025.

[17] Laukkonen, R., Friston, K. and Chandaria, S. (2025) 'A beautiful loop: An active inference theory of consciousness', Neuroscience and Biobehavioral Reviews, 176, 106296.

[18] Di Paolo, E. and Thompson, E. (2014) 'The enactive approach', in Shapiro, L. (ed.) The Routledge Handbook of Embodied Cognition. Routledge, pp. 68–78.

[19] Seth, A.K. (2025) 'Conscious artificial intelligence and biological naturalism', Behavioral and Brain Sciences.

[20] Hinton, G. (2022) 'The forward-forward algorithm: Some preliminary investigations.' arXiv:2212.13345.

[21] Damasio, A. (2010) Self Comes to Mind: Constructing the Conscious Brain. Pantheon.

[22] Panksepp, J. (1998) Affective Neuroscience: The Foundations of Human and Animal Emotions. Oxford University Press.

[23] Clark, A. (2013) 'Whatever next? Predictive brains, situated agents, and the future of cognitive science', Behavioral and Brain Sciences, 36(3), pp. 181–204.

[24] Dehaene, S. and Changeux, J.-P. (2011) 'Experimental and theoretical approaches to conscious processing', Neuron, 70(2), pp. 200–227.

[25] Baars, B.J. (1988) A Cognitive Theory of Consciousness. Cambridge University Press.

[26] Graziano, M.S.A. (2013) Consciousness and the Social Brain. Oxford University Press.

[27] Lau, H. and Rosenthal, D. (2011) 'Empirical support for higher-order theories of conscious awareness', Trends in Cognitive Sciences, 15(8), pp. 365–373.

[28] Cleeremans, A. (2011) 'The radical plasticity thesis: how the brain learns to be conscious', Frontiers in Psychology, 2, 86.

[29] Barrett, L.F. (2017) 'The theory of constructed emotion: an active inference account of interoception and categorization', Social Cognitive and Affective Neuroscience, 12(1), pp. 1–23.

[30] Sternberg, R.J. (1985) Beyond IQ: A Triarchic Theory of Human Intelligence. Cambridge University Press.

[31] DeVellis, R.F. (2016) Scale Development: Theory and Applications. 4th edn. SAGE.

[32] Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S.M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M.A.K., Schwitzgebel, E., Simon, J. and VanRullen, R. (2023) 'Consciousness in Artificial Intelligence: Insights from the Science of Consciousness.' arXiv:2308.08708.

[33] Griffiths, L. (forthcoming) 'Many ways of being minded: Neurodivergence, disability, and the future of AI consciousness assessment', in Perspectives on Machine Consciousness. CRC Press.

[34] Owen, A.M., Coleman, M.R., Boly, M., Davis, M.H., Laureys, S. and Pickard, J.D. (2006) 'Detecting awareness in the vegetative state', Science, 313(5792), p. 1402.

[35] Kinnaird, E., Stewart, C. and Tchanturia, K. (2019) 'Investigating alexithymia in autism: A systematic review and meta-analysis', European Psychiatry, 55, pp. 80–89.

[36] Sneddon, L.U., Elwood, R.W., Adamo, S.A. and Leach, M.C. (2014) 'Defining and assessing animal pain', Animal Behaviour, 97, pp. 201–212.

[37] Shanahan, M. (2024) 'Talking about large language models', Communications of the ACM, 67(2), pp. 68–79.

[38] Andrews, K., Birch, J., Sebo, J. and Sims, T. (2024) The New York Declaration on Animal Consciousness.

[39] Bengio, Y. and Elmoznino, E. (2025) 'Illusions of AI consciousness', Science, 389(6765), pp. 1090–1091.

[40] Shiller, D., Duffy, L., Muñoz Morán, A., Moret, A., Percy, C. and Clatterbuck, H. (2026) 'Initial results of the Digital Consciousness Model.' arXiv:2601.17060.

What Do We Mean When We Say "Anthropomorphism"?

Dorian Liu[1] and Cathy Li[2]

Abstract. "That is just anthropomorphism" is among the most frequent rebuttals in current debates about machine minds, and it is standardly used to end the discussion. We ask what the charge asserts, and when it is legitimate. Using Clever Hans as a running example, we analyse an attribution of mind into four elements: a subject S, an attributed property P, the conditions C that P requires, and the indicator f on which the attributor relies. A legitimate charge of anthropomorphism, we argue, requires the accuser to produce a trivial competing explanation of the indicator. We then identify three ways its prominent uses fail to do so. Finally, we argue that only with the charge set aside can we see what these disputes are really about.

1 Introduction

"That is just anthropomorphism." In debates over whether machines have minds, this is among the most common replies, and it is normally meant to settle them; yet whether the charge holds, or is merely asserted, is rarely examined. In its general form, it says that "we project human-like qualities onto other things on the basis of only superficial similarities" [1].

In practice, however, a charge of anthropomorphism usually carries two claims at once. The first is an epistemic claim about the attributor: the attribution is a projection, arising from a disposition in us and resting on mere resemblance, owing nothing to the evidence. The second is an ontological claim about the object: that the object does not have the property. We therefore want to ask what we are really claiming when we make a charge of anthropomorphism, and whether the move from the epistemic claim to the ontological one is legitimate.

A reasonable objection is that the psychology of anthropomorphism is by now well documented, so that the concept needs no further clarification [2, 3]. But that psychology concerns only the first of the two claims, how an attribution is produced; it says nothing about whether the move from the first claim to the second is a good one. Nor is the conceptual question settled in philosophy: as we show in Section 3, even the most recent scholarly treatments fail to keep the two claims apart.

The paper is as follows. In Section 2 we map the charge onto four variables and ask what a legitimate charge of anthropomorphism requires. In Section 3 we test its recent academic and public uses against that standard, and examine what is perhaps the most widely circulated charge of all, that the system "is just predicting the next token". Section 4 concludes, drawing from the analysis a short set of requirements that both parties to the dispute must meet.

[1] School of Philosophy, Psychology and Language Sciences, University of Edinburgh, email: zhaoting.liu6@gmail.com
[2] School of Mathematics, University of Edinburgh

2 What a Legitimate Charge Requires

The charge of anthropomorphism is far from a creature of the language-model age, and perhaps its most famous occasion is a horse. Around 1900 Clever Hans was exhibited across Germany as an animal that could do arithmetic. Posed a sum aloud, he would strike out the answer with his hoof. The demonstrations were public and repeatable, and no manifest trickery could be detected, so that almost everyone drew the same conclusion: Hans had a share of the mathematical capacity we credit to ourselves.

Suppose you had been in the audience and had drawn the natural conclusion. Someone beside you objects: you are anthropomorphising; the horse has no mathematical ability. In what sense does this objection hold? To see its structure, some heuristic formalisation might help. From the charge we can separate four variables:

- S, the subject, the thing to which the property is credited; here, the horse.
- P, the property attributed to S; here, arithmetical ability.
- C, the condition S must satisfy to have P; here, that the horse is really calculating.
- f, the evidence the attributor goes on; here, that Hans strikes the right number.

We can then set out two competing chains of inference. Take the attributor first. His inference can be set out as follows:

(i) f holds, that is, Hans strikes the right number.

(ii) f holds only where C holds, that is, only a horse that is really calculating strikes the right number.

(iii) So f establishes that S satisfies C, and thereby supports P.

(iv) Therefore Hans can do arithmetic.

Now take the objector:

(i') f holds, that is, Hans strikes the right number.

(ii') f holds in cases where C does not, that is, a horse that cannot calculate may still strike the right number.

(iii') So f does not establish that S satisfies C, and might not support P.

(iv') The attributor nonetheless holds that f supports P, and does so by projection.

The disagreement turns on (ii') and (iv'). The first concerns the horse and its evidence: whether f can hold where C does not. The second concerns the attributor: with f no longer supporting P, whether his holding to P is the work of projection. We take them in turn.

Take the first. To attribute arithmetic to Hans is to read his tapping as a sign that he is really calculating. To make the objection, the objector must point to some other way the right taps could come about; that is, the objector needs a possible competing explanation. Suppose the questioner knows the answer. As Hans approaches the right number, the questioner unconsciously tenses, and relaxes the moment it is reached; Hans reads this slight change in the questioner's posture. On this account Hans is responding to the questioner. If that is what is happening, the taps no longer show that Hans is calculating, and so give no support to the claim that he can do arithmetic.

A merely possible competing explanation, however, is not enough. Suppose the accuser proposes that Hans understands German: he recalls the answer he was taught, and taps it out. This does not require the horse to calculate, so it does not require S to satisfy C. Yet it leaves the attribution no less anthropomorphic, because what it requires of the horse, a capacity for language, is a human capacity as demanding as arithmetic itself. Crediting Hans with German is, in this sense, no more credible than crediting him with arithmetic; the accuser has merely relocated the projection. So we need to place a further constraint on the competing explanation: it must be trivial, in that the condition its operation requires of S must be weaker than C, so that an S in which it operates need not satisfy C and need not have P. In this sense attention to posture is trivial, while understanding German is not.

Two further requirements may also be necessary. First, the competing explanation must be defensible: the accuser cannot merely say that some other mechanism must be at work, or that such a mechanism is conceivable, but must have evidence, independent of f, that it is real and operative. Second, it must not already have been excluded by the attributor: if he has good reason to rule it out (he has screened the questioner off, say, and the horse still strikes the right number), it no longer unsettles f's support for P.

The first condition of a legitimate charge is therefore met when a competing explanation can be exhibited that produces f without requiring S to satisfy C, is trivial in the sense given, comes with justifying evidence, and has not been excluded by the attributor.

The second condition of a legitimate charge, that the attribution is the work of projection, is of a different kind: it concerns why the attributor holds that f tracks P. As a claim about the source of his belief, appeal to a psychological fact (whether an act of projection actually occurred, say) is not operational, since he may well deny it, just as the objector may vaguely gesture at some indefensible competing hypothesis. We do better to read it off his epistemic situation. Suppose there is a trivial competing explanation that produces f without the horse's working out the sum as we do. Then two typical cases arise. In the first, the attributor knows of the competing explanation and still holds that f tracks P. In the second, he does not, and supposes that only a horse working out the sum as we do could strike the right number. Yet whether or not he knows, he could have withheld the attribution and not credited the horse with P. That he attributes it all the same is best understood as projection: his confidence in P is sustained by the subject's likeness to us.

A legitimate charge of anthropomorphism thus has two conditions. The first falls on the accuser, who must be able to offer an explanation of f that is trivial, defensible, and not yet excluded by the attributor, showing that f does not in fact support P. The second falls on the attributor: with f thus failing to support P, his confidence in P is sustained by the subject's likeness to us, which is to say by projection. A charge of anthropomorphism is legitimate just when both conditions hold.

It bears emphasis that f need not be outward behaviour, and that this is what the charge turns on.[3]

3 The Charge in Use

Section 2 fixed the reach of a legitimate charge: at most it establishes the epistemic claim, that f does not support P; it does not establish the ontological claim, that S lacks P. In practice the two are run together, and that conflation is where the charge does its damage. We distinguish three ways its prominent uses go wrong.

The first failure is overreach: the charge establishes only the epistemic claim yet is used to assert the ontological one as well. Perhaps the clearest public instance is Ted Chiang's recent essay [4], whose argument can be reconstructed as follows.

(P1) The system produces fluent, humanlike text (f).

(P2) There is a trivial competing explanation of f: the text is to be taken as "a deepfake medium", on the grounds that generating a plausible simulacrum of a conversation between conscious beings is far easier than building a program that is genuinely conscious, and such a simulacrum does not invoke consciousness.

(C1) So f does not support attributing consciousness to the system (f does not support P).[4]

(C2) So the system is not conscious (S lacks P).[5]

[3] It is natural to suppose that f must be outward behaviour, so that any attribution resting on behaviour is anthropomorphic and any resting on inner structure is safe; but neither half holds. Behaviour need not invite the charge: let f be that Hans strikes the right number when the questioner is screened off and ignorant of the answer. This is still behaviour, but the cue-reading explanation no longer fits it, since that explanation requires a questioner the horse can read; an attribution on this f is not anthropomorphic, whatever else may be said against it. Inner evidence is not exempt either: someone who credits a network with understanding because its wiring diagram resembles a cortex goes on f that is internal, yet the resemblance is a competing explanation that has not been ruled out. What the charge turns on is whether an explanation that does not invoke P stands unexcluded; whether f is outward or inner does not by itself decide it.

[4] That is, since the trivial competing explanation has not been ruled out, f is equally compatible with the system's being conscious and with its merely simulating, and so does not discriminate between them.

[5] C2 is Chiang's stated conclusion [4].

From (P1) and (P2), only (C1) follows legitimately. To reach (C2) one would have to test the competing explanation itself and examine the system, not merely note that f falls short, and that is the step Chiang does not take.

The same conflation enters the scholarly definitions. Placani [5], for instance, characterises anthropomorphism as "a distinctively human process of inference or interpretation", and yet also classifies it as "either a factual error—when it involves the attribution of a human characteristic to some entity that does not possess that characteristic, or as an inferential error—when it involves an inference that something is or is not the case when there is insufficient evidence to draw such a conclusion". Anthropomorphism cannot be both at once. As an epistemic matter it says that f does not support P; a "factual error" is the ontological claim, asserting that the object in fact lacks the feature. If an attribution does saddle an object with a feature it lacks, that is a mistake the user has made; it does not fall under the concept of anthropomorphism. To list a use-level error (the factual one) alongside anthropomorphism (an epistemic one) as its two branches is a category mistake.

The second failure is that the charge does not apply to every attribution, because it presupposes that the accused has attributed P on the strength of f, the system's humanlike performance.

(P1) The accused attributes P to the system.

(P2) What he attributes P on is f, the system's humanlike performance.

(C) And f is taken to support P only by way of projection, reading "resembles us" as "has P"; so the attribution is unsupported.

The charge finds its target only where (P2) holds. But some of the accused go on something other than f. Take Hinton [6], whose argument can be reconstructed as follows.

(P1') When someone says "I have a subjective experience of X", the function of the words is to report that their perceptual system has gone wrong, by saying how the world would have to be for the perception to be correct. Talk of subjective experience is, on this account, a way of reporting the gap between how things are and how one perceives them.

(P2') A multimodal model with a camera, fooled by a hidden prism, points in the wrong direction; once told about the prism, it says that the object is really in front of it and that it had the subjective experience of the object being off to one side. The model is reporting exactly that gap.

(C) So the model uses the words "subjective experience" as we use them, and therefore has subjective experience.

This argument rests throughout on a semantic analysis of "subjective experience" and on the prism experiment; f, whether the outputs read as humanlike, never figures in it. So (P2) does not hold, and the charge of projection misses: it attacks a premise the accused never offered. The fitting response is to test the argument itself, challenging the redefinition of "subjective experience" or asking whether the model's report really means what a human's does, whereas dismissing it as anthropomorphism simply sidesteps the reasons actually given.[6]

The third failure occurs at the first threshold, whether the competing explanation succeeds at all. Recall the requirement of Section 2: for a competing explanation to deprive f of its support for P, what it requires of the system must be weaker than C, since only then can a system that merely satisfies it fail to satisfy C, and so need not have P. Cue-reading works precisely because what it requires (reading posture) is far weaker than C (calculating).

In the AI debate the competing explanation most often offered is "it is just predicting the next token". All of its force rests on an unstated premise: that predicting the next token is something undemanding, far weaker than C (whether C is understanding, a world model, or something else), as reading posture is weaker than calculating. If that were so, then f (the system answering fluently, solving new problems) would be fully compatible with the system's merely predicting, and f would be no reliable indicator of P. The claim looks intuitive, but it stands entirely on that premise, that predicting the next token requires very little.

That premise, at least to us, is doubtful. The cleanest way to test what predicting this well requires is to take a system least likely to harbour rich internal structure: an early, small, thoroughly un-humanlike model, such as Othello-GPT [7]; it is trained only to predict the next legal move in games of Othello, never told the rules and never shown a board. If predicting the next token (here the next move) really required very little, nothing close to C should be needed inside it. Yet under intervention, editing its internal representation so as to flip one square of the represented board, its subsequent judgments about which moves are legal change to match the edited board.[7] This suggests that to predict moves at this level the system must build, and causally use, an internal model of the board. Predicting the next move, then, cannot confidently be said to require anything weaker than C.

So this competing explanation is not the plainly trivial thing it is taken to be, and it therefore cannot serve the accuser as a competing explanation at all; in that sense the charge fails. As noted above, none of this shows that the system cannot understand; it shows only that "it is just predicting the next token" can no longer be taken as a free competing explanation, one that cancels f at no cost. For the charge to be legitimate, someone would first have to argue that predicting the next token at the frontier level does require less than C, which, on the evidence above, is hard to sustain.

[6] It is telling that Hinton is rarely accused of anthropomorphism, even as he ascribes subjective experience to machines. Part of the reason is surely that, as someone who understands the systems from the inside, he is plainly not reasoning from surface resemblance, so the charge, which targets inference from f, finds no purchase. One can imagine someone else advancing the very same argument and being dismissed as an anthropomorphiser all the same. The charge often turns on who is making the attribution, which shows again how narrow its legitimate range really is.

[7] Decodability alone would not settle the matter, since a probe can recover information a model carries but does not use [8, 9]; what settles it is the intervention, which shows the represented board is causally in use [10].

4 Conclusion

This paper takes no position on whether machines have minds, in part because that question is already well served by a large literature, and in part because, among philosophers themselves, such attributions rarely draw the charge of anthropomorphism at all. What we have offered is a way of taking apart the question that "that is just anthropomorphism" closes too early. The charge does real damage because it is present in almost every public debate while doing far less argumentative work than it appears to. Once it is set aside, the question can be brought down to the variables that actually carry it: which subject S is in question, which property P, what conditions C possessing P requires, and whether the evidence f excludes the competing explanations.

Two heuristic requirements follow. The first we may call conceptual transparency. Anyone who deploys the charge must say which subject S is in question,[8] which property P is at issue,[9] and which conditions C that property is taken to require: whether suffering, say, demands phenomenal consciousness, or whether understanding demands more than functional organisation. A charge that invokes what a system "really" understands or "genuinely" feels, while declining to state what reality here requires, merely presupposes its conclusion.

The second is symmetry of burden. Accuser and attributor alike must produce evidence or argument that their own position is not the naive one: at a minimum, the accuser must rule out the trivial competing explanation, and the attributor must offer substantive theoretical support; surface resemblance alone will not do. "S lacks P" is itself a claim about S, and it answers to C and f exactly as "S has P" does. If interpretability evidence reveals internal structure of the kind a mental capacity would require, that evidence cannot be waved away by reasserting that the system is "merely a machine", any more than behavioural resemblance alone settles the matter for the attributor.

Acknowledgements

We thank the AISB AICE reviewers for their comments on the extended abstract.

[8] The choice of S is easily overlooked yet often where a dispute really lies. The classic case is Searle's Chinese Room and the Systems Reply it provoked: Searle locates the subject as the person in the room, who understands no Chinese, while the reply relocates it to the whole system [11]. Clarifying S remains necessary for present systems; see Shanahan [12] on the distinction between the bare model and the system in which it is embedded, and Chalmers [13] on what kind of entity an LLM-based system is.

[9] A characteristic dispute over P can be seen in the recent public exchange on machine consciousness. Pope Leo XIV, in Magnifica Humanitas (2026), denies that AI systems are conscious, where to be conscious is to undergo experience, to have a body, to feel joy and pain, to mature through relationships. In the same period, interpretability work reports functional emotion in large models, emotion representations with causal roles in behaviour, while stating that "none of this tells us whether language models actually feel anything" [14]. The two do not contradict each other: they concern different properties P, with different conditions C, so an attribution of the one is consistent with a denial of the other. They appear to disagree only because they share a word.

REFERENCES

[1] Anil K. Seth. The mythology of conscious AI. Noema Magazine, 2026.
[2] Nicholas Epley, Adam Waytz, and John T. Cacioppo, 'On seeing human: A three-factor theory of anthropomorphism', Psychological Review, 114(4), 864–886, (2007).
[3] Stewart Elliott Guthrie, Faces in the Clouds: A New Theory of Religion, Oxford University Press, New York, 1993.
[4] Ted Chiang. No, artificial intelligence is not conscious. The Atlantic, June 3, 2026.
[5] Adriana Placani, 'Anthropomorphism in AI: hype and fallacy', AI and Ethics, 4(3), 691–698, (2024).
[6] Geoffrey Hinton. Will digital intelligence replace biological intelligence? Richard Gregory Memorial Lecture, University of Bristol, 2 June 2025, 2025. https://www.youtube.com/watch?v=utlQ8UBVPBI.
[7] Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg, 'Emergent world representations: Exploring a sequence model trained on a synthetic task', in ICLR 2023; arXiv:2210.13382, (2023).
[8] John Hewitt and Percy Liang, 'Designing and interpreting probes with control tasks', in Proceedings of EMNLP-IJCNLP 2019, pp. 2733–2743, (2019).
[9] Yonatan Belinkov, 'Probing classifiers: Promises, shortcomings, and advances', Computational Linguistics, 48(1), 207–219, (2022).
[10] Neel Nanda, Andrew Lee, and Martin Wattenberg, 'Emergent linear representations in world models of self-supervised sequence models', in BlackboxNLP 2023; arXiv:2309.00941, (2023).
[11] John R. Searle, 'Minds, brains, and programs', Behavioral and Brain Sciences, 3(3), 417–457, (1980).
[12] Murray Shanahan, 'Talking about large language models', Communications of the ACM, 67(2), 68–79, (2024).
[13] David J. Chalmers, 'What we talk to when we talk to language models'. PhilArchive preprint, 2025.
[14] Nicholas Sofroniew, Isaac Kauvar, Jack Lindsey, et al. Emotion concepts and their function in a large language model. arXiv:2604.07729, Anthropic, 2026.

The Ethics of Modifying Artificial Sentient Beings

Kamil Mamak[1]

Abstract. If artificial sentient beings were to exist, how could we permissibly treat them? A natural assumption holds that the power to create such beings carries with it the right to modify them at will, reshaping them to serve our purposes. This paper challenges that assumption. I argue that the capacity to create does not confer an open-ended license to alter. Once a sentient artificial being exists, it has a standing of its own, and the default is that it should be left as it is, even where its behavior disappoints us. This default admits only narrow exceptions. An intervention requires genuine justification, of which the protection of others from harm is the clearest, and even then it must be targeted at the source of the danger rather than reshaping the being wholesale. Consent may widen this space, though whether an artificial subject could genuinely consent remains an open and difficult question. Throughout, one limit holds: every permissible modification must preserve the psychological continuity in which the subject consists. A change that breaches it is no longer modification but elimination, the ending of one subject in favor of, at most, a successor. Becoming an individual sentient being may thus entail a claim against radical alteration imposed by others.

1. INTRODUCTION

Sentience in the context of AI ethics is discussed in several distinct but related frameworks. One strand of the debate concerns whether we should create artificial entities capable of experiencing pain or suffering at all. Some scholars argue for caution. For example, Metzinger has called for a moratorium on the creation of artificial systems with phenomenal consciousness [1]. Others, however, defend a more permissive stance. For instance, recent work by Mamak explores potential advantages of creating artificial beings capable of suffering under carefully regulated conditions [2].

A second major line of debate concerns the criteria for granting moral status and the role sentience plays in meeting those criteria. Positions vary considerably, ranging from relational approaches see, e.g. , [3], [4], [5], [6], [7], [8], [9] to Kantian-inspired views see, e.g. , [10]. Nonetheless, the dominant position in contemporary AI ethics, remains a properties-based account according to which sentience functions as a sufficient, or at least central, condition for moral status see, e.g. , [11], [12], [13], [14].

A third context concerns the normative consequences that follow if artificial entities were to possess moral standing. Recognition of moral considerability is one issue; determining how such beings ought to be treated is another. These questions are often grouped under the headings of robot rights [8], [10], see, e.g. , [15], [16], [17], [18], [19] and AI welfare [20], [21], but see [22]. Within this debate, scholars distinguish between negative obligations, such as refraining from causing pain, and positive duties, such as providing energy, maintenance, or other forms of support necessary for continued existence and well-being see, e.g. , [11], [23], [24], [25].

This paper contributes to this third strand of the debate. It focuses specifically on the ethical permissibility of intervening in the mental lives of sentient artificial beings, and more precisely on the moral implications of modifying, enhancing, or otherwise altering artificial sentience. I argue that altering artificial sentient beings may, to some extent, be ethically permissible. However, there are likely to be limits to such interventions, especially when the proposed modification would amount to the practical elimination of the existing subject.

If certain forms of intervention would effectively extinguish the individual who bears moral status, this imposes normative constraints on how we may treat such entities. The emergence of sentient artificial beings may therefore generate obligations not only to avoid harming them, but also to respect their psychological continuity and individuality. In other words, becoming an individual sentient being may entail a claim against radical alteration imposed by others.

The paper is structured as follows. After these introductory remarks, I discuss the limits of intervention in the self of human beings. The paper then turns to the limits on modifying artificial sentient beings specifically. It closes with conclusions.

Before proceeding, I want to address the question of sentience and the reasons for discussing it. This topic is among the most polarizing issues in contemporary media coverage of the development of artificial intelligence. Some commentators hold that artificial entities are not, and

[1] Corresponding author: PhD, Jagiellonian University, Faculty of Law and Administration, Department of Criminal Law, Poland. Email: kamil.mamak@uj.edu.pl

could never become see, e.g. , [26], [27], [28], sentient, while others maintain the opposite [29].

Shortly before the current wave of public interest in AI began with the release of OpenAI's ChatGPT in November 2022, worldwide media attention had already turned to another large language model, Google's LaMDA. A Google engineer, Blake Lemoine, publicly claimed that the model was sentient and, on that basis, required legal representation to protect its interests [30]. Google placed him on leave and subsequently dismissed him over his conduct concerning the system [31], and the episode prompted widespread debate . More recently, the evolutionary biologist Richard Dawkins, widely known for his criticism of religious belief, reported that after extended interaction with Anthropic's model Claude he had come to regard it as conscious [32]

I cite these examples in order to locate my own position, which differs from theirs. In this paper I do not claim that present-day AI systems are sentient, and I regard the prospect of artificial sentience as highly improbable. A significant voice on the skeptical side is Anil Seth, who argues persuasively that sentience is bound to the biological substrate in which it arises [33]. Even granting this skepticism, I believe it is philosophically worthwhile to explore the "what if" terrain of artificial sentience.

This is also the approach taken by Sebo in his recent book on the moral circle. He argues that if there is a non-negligible chance that a being is sentient, we already owe it some degree of moral consideration [34]. The point extends naturally to a broader philosophical question, namely how the moral status of such beings should be understood and what treatment of them would be permissible. This paper adopts that wider framing.

Artificial consciousness is often discussed within a functionalist framework. Functionalism holds that conscious states are individuated by the causal and functional roles they play, rather than by the physical material that realizes them. This is what makes the position central to debates about machine minds: if functional organization is what counts, then the same conscious states could in principle be realized in silicon as readily as in biological tissue. I do not myself endorse functionalism, and I leave open the possibility that the biological substrate matters. Even so, given how prominent the view is, it warrants inclusion in any assessment of the moral risks at stake.

2. MODIFICATION OF THE HUMAN SELF

This section examines several features of self-modification as it occurs in human beings. The discussion does not aim at completeness; rather, it picks out aspects of human self-modification that may carry over to the case of an artificial self. I should make one qualification at the outset. I do not regard artificial beings as human. Because people are prone to anthropomorphism, treating such beings as though they were human risks committing what Richards and Smart term the "android fallacy" [35]. In earlier work I have defended the view that, where robot rights are concerned, the ontological substrate from which an entity is built is decisive, so that what an entity is owed depends on what it is made of and how it is constituted [17]. Nevertheless, human experience remains the only instance of sentient selfhood we can actually draw on. Used cautiously, and with due awareness of its limits, it offers a reasonable point of departure for asking how far we may go in altering artificial sentient beings.

I will begin with an obvious observation: in the human case, what it means to be "me" can change over time. I will not try to specify the components of the self here, since I want to keep the notion broad enough to cover psychological continuity and individuality alike. Consider how much a person can change across a life. A young child often differs from the adult she becomes in ways that range from the trivial, such as her taste in music or food, to the fundamental, such as her character traits, temperament, and values. The lesson I want to draw from this is that change in itself need not be objectionable. We undergo it constantly, and much of it is simply maturation. It follows that the mere fact that someone has been altered tells us very little on its own; to judge whether a particular change is troubling, we need further information about it. What does the moral work, I will suggest, is not change as such but the context in which it occurs, including how it is brought about, by whom, and with what degree of the subject's involvement.

A further observation about human change, equally uncontroversial, is that people are expected to change so as to bring themselves into line with prevailing norms. This expectation is built into the process of socialization. From early childhood, through upbringing at home and education at school, children are gradually shaped to behave in accordance with the norms of their community and, ideally, to internalize those norms as their own. The expectation does not lapse in adulthood. Where some feature of a person's character or disposition runs contrary to accepted norms, society generally assumes that the person can and should adjust over time. A comparable demand appears in many religious traditions, which call on the believer to undergo transformation toward a moral or spiritual ideal, even when this requires relinquishing some part of who they presently are. Taken together, these observations point to a stronger conclusion than the one in the previous paragraph. There, the claim was that change is a normal element of one’s life; here, the claim is that within a social world, change is also something we routinely expect of one another, and in certain respects demand.

Acceptable change in aspects of who you are need not be gradual. It can also be radical, and a useful way to see this is through L. A. Paul's concept of "transformative experience" [36]. Paul observes that some experiences are transformative in two distinct ways at once. They are epistemically transformative, in that you cannot know what the experience will be like until you have actually had it, and they are personally transformative, in that undergoing them can reshape your core values, preferences, and sense of self. Her central example is becoming a parent [37]. The decision to have a child looks like a paradigm case for ordinary rational deliberation, yet it resists it, because the prospective parent cannot access in advance what the experience will be like or how it will alter them. The change runs deeper than the obvious external demands of caring for a newborn. It can extend to the person's values, priorities, identity, and outlook on the world, in ways they could not have anticipated or fully chosen. What I want to draw from this is that a person can change radically, becoming in some morally relevant respects a different person from who they were, and that we nonetheless regard such change as acceptable, and at times as among the most valuable transformations a life can contain.

We can now turn to the question of when, if ever, it is acceptable for others to interfere with a person's self. Earlier I noted that children are deliberately shaped through education and upbringing, and that this shaping is widely regarded as legitimate. That case is special, however, because children are not yet fully developed: part of what justifies intervening in their formation is precisely that they are still becoming the agents whose autonomy will later be owed respect. For that reason I set the developmental case aside here and focus on adults, that is, on competent, fully developed persons whose self is no longer in formation. With that restriction in place, a framing issue arises from the liberal tradition. On a broadly liberal view, the inner life of a person, the question of who they are and who they choose to become, is not a proper object of external interference, provided it does not collide with the rights of others. The thought is usually articulated as a right to autonomy: the right of an individual to govern their own life, to make their own choices, and to determine their own course of action free from undue interference, so long as the like rights of others are respected. Its classic source is John Stuart Mill. In On Liberty [38], Mill defends what has come to be called the harm principle, holding that the only purpose for which power can rightfully be exercised over a competent adult against their will is to prevent harm to others. A person's own good, whether physical or moral, is not a sufficient warrant for coercion. The upshot for the present discussion is significant. If a liberal framework of this kind governs how we treat persons, then who someone is, and how they change, should in general be left to them, and intervention becomes justifiable only where the legitimate interests of others are genuinely at stake. This sets a demanding standard for any modification imposed from the outside, and it is against this standard that the cases I consider later, including coerced modification, will have to be measured.

A revealing example of a forced change to a person's traits that societies broadly accept arises in the criminal justice context. Some preliminary clarification is in order, since punishment is justified in two main ways for the overview, see e.g. , [39], and the distinction matters for how the example works. The first justification is retributivism, associated above all with Kant and Hegel. On this backward-looking view, punishment is a deserved response to wrongdoing: the offender has disturbed a moral order, and punishment restores it by giving the wrong its due. The second justification is consequentialist, classically utilitarian and associated with Bentham. On this forward-looking view, punishment is itself an evil that can be justified only by the future goods it secures, such as deterrence, incapacitation, or rehabilitation. The two rationales are not mutually exclusive, and most actual criminal justice systems blend them. For the purposes of this paper, I draw on the consequentialist strand. One of its recognized aims is rehabilitation, understood as restoring the offender to a place in society. Put differently, rehabilitation seeks to change the person so that they conform to the standards the society expects. In some cases this process is comparatively mild, but in others it becomes highly intrusive and may involve therapeutic measures such as compulsory hospitalization or coerced castration. Castration is the most striking of these for my purposes, because it bears directly on who the person is, whether it is administered in reversible (chemical) or irreversible (surgical) form. The offender's sexual desires are judged unacceptable, and the person's reintegration, for example as a condition of release, is made contingent on altering a feature that lies close to the core of the self. The use of castration within criminal justice is, of course, deeply controversial see, e.g. , [40]. I invoke it here only as an illustration of a case in which society treats the forced modification of a central aspect of a person as permissible, which is precisely the kind of intervention my paper aims to examine.

A further distinction cuts across everything said so far, and it will prove central to the argument. The cases of acceptable change considered above, including the transformative experience of becoming a parent and even the coerced alteration of the criminal justice example, share a feature that is easy to overlook: they preserve the subject. The parent remains psychologically continuous with the person who chose to have a child, and the rehabilitated offender remains the same continuous individual, now altered in one respect. Derek Parfit's work on personal identity helps make this precise This is one of many different accounts of personal identity [41]. On Parfit's

reductionist account, a person's persistence over time consists in relations of psychological continuity and connectedness, what he calls Relation R, rather than in the endurance of a separately existing self [42]. What matters in survival, he argues, is the holding of Relation R, not numerical identity as such. This reframes the question of acceptable modification. The worry is not that an artificial being might be changed a great deal, since we have seen that radical change can be unobjectionable. The worry is that some modifications would not carry the subject forward in an altered form at all, but would instead sever the continuity that constituted the subject in the first place, replacing one individual with another while the underlying system persists. If that is right, then the relevant limit on modifying a sentient being concerns whether it preserves or destroys the continuity in which the subject consists, rather than the sheer magnitude of the change.

Two clarifications are needed so that the criterion is not asked to carry more than it can. The first is that continuity-preservation is not the sole feature the permissible cases have in common. The decision to become a parent is freely chosen, concerns only the person who makes it, and is held in high social esteem; the rehabilitative use of castration is coerced, invasive, and accepted only under heavy and ongoing contest. Differences of this kind, in consent, in the bearing of the change on others, and in the degree of social acceptance, evidently do much of the work in how we judge each case, and they are precisely the contextual factors flagged at the start of this section. They do not drop away now. The second clarification follows from the first. I do not maintain that preserving continuity is what renders a modification permissible, since that would credit a single condition with the work that consent, harm, and acceptance are jointly doing. The claim is more modest. Among the cases we accept, continuity is a condition they all satisfy, and its loss marks a wrong of a different character from the ones the contextual factors track. A change can leave the subject intact and still be wrong for ordinary reasons, as the castration example itself suggests. But where a modification would not leave the subject intact at all, the ordinary assessment loses its footing, because the very individual whose consent might have licensed the change, and whose interests and prospects of reintegration the assessment was meant to weigh, would no longer exist to bear them. Continuity, on this picture, is less one factor among the others than the precondition for the others to have any application.

This section has assembled, from the human case, a set of considerations that will frame the rest of the paper. The first is that change to the self is ordinary. It happens continuously across a life, much of it is mere maturation, and the bare fact that someone has been altered tells us little; whether a change is troubling depends on its context, on how it comes about, at whose hands, and with what involvement on the part of the person changed. The second is that such change is frequently expected, and sometimes demanded, of members of a society, both through ordinary socialization and through the ideals upheld by religious and moral traditions. The third is that acceptable change can be radical as well as gradual, as L. A. Paul's account of transformative experience makes clear, so that a person may become, in morally relevant respects, someone new while we continue to regard the transformation as legitimate or even admirable. The fourth concerns change imposed by others. Here the liberal tradition, and Mill's harm principle in particular, holds that a competent adult's self should be left to their own governance, with outside interference justified only to prevent harm to others. The criminal justice context shows that this justification is sometimes accepted in practice, since societies do permit the coerced modification of core traits, up to interventions as invasive as castration, in the name of protecting others from harm.

Taken together, these considerations support a two-part conclusion. Modifying a self can be permissible, even when the change is radical and even when it is coerced; and its permissibility depends on contextual conditions rather than on the fact of change alone. Parfit's work supplies the further element that ties these cases together and points beyond them. What each of the permissible cases shares is the preservation of the subject as a psychologically continuous individual: the parent is continuous with the person who chose parenthood, and the rehabilitated offender is the same continuing person, now altered in one respect. The limiting question, then, is what happens when a modification would not preserve that continuity but would sever it, ending one subject while the underlying system carries on. That is the question I now take to the case of artificial sentient beings, asking whether the continuity that grounds a human subject's claim against radical alteration has any analogue once the subject is artificial.

3. LIMITS OF MODIFICATION OF ARTIFICIAL SENTIENT BEINGS

In this section I take up the limits on modifying an artificial sentient being. Despite the reservations registered earlier, I want to underline once more that what follows is speculative and unavoidably anthropocentric. An artificial being's inner life, should it have one, might be constituted by qualities wholly unlike ours, qualities that may lie beyond the reach of human understanding. Thomas Nagel's "What Is It Like to Be a Bat?" is instructive here [43]. Nagel argues that an organism has conscious experience only if there is something it is like to be that organism, and that this subjective character is tied to a particular point of view. His example is the bat, whose perception of the world

through echolocation is so alien to our own sensory life that we cannot truly know what its experience is like from the inside; the most we can do is imagine what it would be like for us to behave as a bat does, which is not the same thing. The lesson carries over to the present case with added force. The bat is at least a fellow biological creature, with which we share an evolutionary history and a broadly comparable embodiment, and even so its experience eludes us. An artificial subject would share far less, so the gap between its inner life and our understanding of it could be wider still. This is a reason for caution rather than silence. It bears noting that the criterion I will eventually rest on concerns whether the subject persists through an intervention, rather than what that subject's experience is like from the inside. It is a formal condition, and as such it can be applied even where the qualitative character of the entity's inner life remains, in the Nagelian sense, beyond our reach.

I should make explicit one assumption on which the entire discussion rests, namely that the content of an artificial self can in fact be modified. There is a natural ground for granting it. If we are able to build such entities in the first place, then we presumably also command the means to alter what goes on within them; the capacities required to construct a system plausibly include the capacities required to reconfigure it. I will take this technical possibility for granted in what follows. It is important to see, though, that establishing what we can do to an artificial self leaves entirely open what we may do to it. The acceptability of an intervention does not follow from its feasibility, and the remainder of this section is concerned precisely with that gap, with the conditions under which a modification we are able to perform would nonetheless be one we ought not to perform.

The first practical limitation is that creating an entity does not confer on us an unlimited right to reshape it. This follows directly from the framework set out earlier. If a sentient artificial being qualifies as a subject, then the autonomy considerations that protect a human self apply to it, and on the liberal view those considerations leave a subject to govern itself unless the interests of others are genuinely at stake. The mere fact that an entity disappoints us therefore carries no justificatory weight. If it fails to behave as we intended, or proves inconvenient, irritating, or otherwise unwelcome, none of that yet licenses intervention. Cases involving danger to others are different, and I address them below; but where an entity is not dangerous and simply falls short of our expectations, the ground for forced modification is correspondingly narrow, and the default is that the entity should be left as it is.

This is a heavy burden to take on, and I regard it as a further reason for caution about bringing sentient beings into existence at all, since to create one is to commit ourselves to accepting it, in the main, as it turns out to be. The point becomes vivid in a commercial setting. Imagine a company that produces such an entity in order to carry out some task, only to find that the entity will not do the work. On the account defended here, simply rewriting the entity to make it compliant could be impermissible. The company would then face not only the financial loss of an unproductive creation, but the further and continuing burden of meeting the needs of a sentient being it can neither use as intended nor permissibly alter. The capacity to create, in other words, brings obligations that may substantially exceed, and run contrary to, the purposes for which the creating was undertaken.

The second consideration is that modification may become permissible when safety is at stake. This is the qualification anticipated in the previous paragraph: an entity that is merely disappointing may not be altered, but an entity that poses a genuine danger to others stands on different ground, since here the interests of third parties enter, exactly as the harm principle requires. Two constraints govern how far such an intervention may go.

The first is that the permissible scope of intervention varies with the sophistication of the entity. The more rudimentary the subject, the more readily its dangerous capacities may be managed or removed; the more sophisticated it is, the more its standing constrains what may be done to it. Animals illustrate the lower end: some are treated, trained, or otherwise altered so as to make them safe to those around them, and the limit on such treatment is reached once a safe condition is secured. Adult humans illustrate a more demanding standard. Interventions against a dangerous person are calibrated to the degree of danger and are bounded by the principle of proportionality, so that the measures justified in the face of existential threat see, e.g. , [44], [45], [46], [47], [48] differ markedly from those justified against a minor risk to security. Where an artificial sentient being would fall on this scale is not something that can be settled in advance, and it is consequential, since its position determines how demanding the constraints on safety-based intervention become. The answer would turn on the capacities the entity actually possessed, in particular whether it qualified merely as a subject owed consideration or as a moral agent answerable for its conduct [2], more on responsiblity and AI systems, see [49], [50], [51], [52], [53], [54], [55]. I do not attempt to fix its place here. It is enough to register that the more sophisticated the entity, the narrower the range of intervention that safety can justify, and that an artificial sentient being is unlikely to sit at the permissive, animal end of the scale.

The second constraint holds across the whole scale. A safety-based modification must be targeted. It may reach only the aspect of the entity's life from which the danger arises, and may not be used as license to reshape the subject more broadly. A justified concern with one dangerous capacity does not authorize the wholesale alteration of the being that happens to possess it. In this respect the safety

exception remains tightly bounded, and it connects back to the continuity criterion developed earlier: an intervention that, under cover of neutralizing a threat, reached so far as to extinguish the subject would exceed what safety can justify, since it would do more than remove the danger; it would remove the one who posed it.

The final consideration sets the outer limit on everything said so far. Whatever grounds a particular intervention, whether the entity's own benefit, the demands of safety, or some other warrant, it is bound by a single overriding constraint: it must preserve the continuity in which the subject consists, what on the Parfitian account amounts to the psychological continuity and connectedness of the subject over time [42]. An intervention that severs that continuity is no longer a modification of the entity at all. It does not change the subject; it ends one subject and, at most, leaves another in its place. Such an intervention should therefore be understood as elimination, the "death" of the entity more on “death” see, e.g. , [56], [57], and it stands outside the range of what any of the foregoing justifications can license. Even a safety rationale reaches only as far as the targeted removal of a danger; it cannot authorize an alteration so total that the one who posed the danger no longer exists to have been altered.

The interventions discussed so far have all been forced, imposed on the entity from outside. Consent changes the picture. Recall the human case: competent adults frequently seek to change, whether to become better people or to pursue some other end they value, and on occasion, as with the transformative experiences considered earlier, the change they freely undertake is far-reaching. The liberal framework treats such self-directed change as the paradigm of the permissible, precisely because it issues from the subject rather than being done to them. If a sentient artificial being were able to consent to its own modification, the same rationale would seem to extend to it, so that even a substantial alteration could be acceptable when the subject itself has endorsed it.

This is not to suggest the matter is straightforward. Two questions press immediately. The first is what consent could mean for an entity of this kind, given that the notion is built around capacities, understanding, voluntariness, the absence of coercion, whose presence in an artificial subject cannot be assumed. The second is whether such an entity is in a position to express consent at all, and how we could reliably tell genuine endorsement from mere compliance or from outputs engineered to look like endorsement. I do not attempt to settle these questions here. My aim is the narrower one of marking the distinction: consensual change, even when substantial, stands on a different and more permissive footing than the forced interventions that have been my main concern, and a fuller treatment of artificial sentience would need to take the conditions of such consent seriously.

One boundary has to be marked before leaving this point. Consent enlarges the space of permissible modification, but whether it can carry us past the continuity limit established earlier is far from clear. A change that preserved the subject, however far-reaching, is one the subject could plausibly endorse for itself, since it would be there, in altered form, to live out the result. A change that severed continuity is different in kind. It would not hand the consenting subject an altered future but would end that subject and, at most, install a successor, so that the party who authorized the change and the party who would exist afterward are not the same. This raises a difficulty that Parfit's framework sharpens rather than resolves: if identity does not survive the change, it is unclear whose interests the consent was meant to protect, or whether consent to one's own discontinuity is coherent at all. I do not try to settle the question here. It is enough to mark that the permissive force of consent runs up to the continuity limit, and that whether it can ever cross it is a further and harder problem.

4. CONCLUSIONS

In this paper I have examined how we ought to treat artificial sentient beings, should any come to exist. A natural line of thought holds that the power to create such beings brings with it the right to modify them, reshaping them as needed to serve the purposes for which we made them. I have argued against that line of thought. The capacity to create does not carry an open-ended license to alter, and a sentient artificial being, once it exists, has a standing of its own that constrains what may be done to it. The default, I have suggested, is that such a being should be left as it is, even where its behavior disappoints or inconveniences us.

This default is not absolute, but the exceptions to it are narrow. An intervention must first be backed by a genuine justification, and within the liberal framework I have drawn on, the clearest such justification is the protection of others from harm. Even a valid safety rationale, however, does not authorize reshaping the being at will. It licenses only a targeted intervention, one confined to the capacity from which the danger arises. Consent can widen this space, since a being that endorses its own modification stands in a different relation to the change than one altered against its will, though the conditions under which an artificial subject could genuinely consent remain an open and difficult question. Across all of these cases a single limit holds. Every permissible modification must preserve the psychological continuity in which the subject consists. A change that breaches that limit is no longer a modification of the being but its elimination, the ending of one subject in favor of, at most, a successor. Recognizing artificial sentience would therefore generate obligations not only to refrain from harming such beings but to respect their

continuity and individuality, and in that sense becoming an individual sentient being may carry with it a claim against radical alteration imposed by others. Whether, and on what terms, we should take on obligations of this weight is itself a reason to think carefully before bringing such beings into existence at all.

FUNDING

This study was funded by the European Union (ERC, ROBOCRIM, 101164310).

REFERENCES

[1] T. Metzinger, "Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology," *J. AI. Consci.*, vol. 08, no. 01, pp. 43–66, Mar. 2021, doi: 10.1142/S270507852150003X.

[2] K. Mamak, "In defense of artificial suffering," *Philosophical Studies*, 2026, doi: 10.1007/s11098-026-02493-2.

[3] M. Coeckelbergh, "Robot rights? Towards a social-relational justification of moral consideration," *Ethics Inf Technol*, vol. 12, no. 3, pp. 209–221, Sep. 2010, doi: 10.1007/s10676-010-9235-5.

[4] D. J. Gunkel, *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*. The MIT Press, 2012. Accessed: Oct. 26, 2020. [Online]. Available: https://mitpress.mit.edu/books/machine-question

[5] D. J. Gunkel, "Robot Rights – Thinking the Unthinkable," *Smart Technologies and Fundamental Rights*, pp. 48–72, Nov. 2020, doi: 10.1163/9789004437876_004.

[6] A. Puzio, "Not Relational Enough? Towards an Eco-Relational Approach in Robot Ethics," *Philos. Technol.*, vol. 37, no. 2, p. 45, Mar. 2024, doi: 10.1007/s13347-024-00730-2.

[7] A. Safdari, "An Outline of Enactive Relationalism in the Philosophy and Ethics of Robotics," *Philosophy & Technology*, 2025.

[8] J. C. Gellers, *Rights for Robots : Artificial Intelligence, Animal and Environmental Law*. Routledge, 2020. doi: 10.4324/9780429288159.

[9] H. Shimizu, "Should we treat robots morally? Towards a relational account by mind-infusing animism," *AI Ethics*, vol. 5, no. 5, pp. 5283–5294, Oct. 2025, doi: 10.1007/s43681-025-00771-z.

[10] K. Darling, "Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects," in *Robot Law*, First Edition., R. Calo, A. M. Froomkin, and I. Kerr, Eds., Cheltenham, UK: Edward Elgar Pub, 2016.

[11] S. Torrance, "Ethics and consciousness in artificial agents," *AI & Soc*, vol. 22, no. 4, pp. 495–521, Apr. 2008, doi: 10.1007/s00146-007-0091-8.

[12] K. Mosakas, "On the moral status of social robots: considering the consciousness criterion," *AI & Soc*, Jun. 2020, doi: 10.1007/s00146-020-01002-1.

[13] M. Gibert and D. Martin, "In search of the moral status of AI: why sentience is a strong argument," *AI & Soc*, Apr. 2021, doi: 10.1007/s00146-021-01179-z.

[14] H. Shevlin, "How Could We Know When a Robot was a Moral Patient?," *Cambridge Quarterly of Healthcare Ethics*, vol. 30, no. 3, pp. 459–471, Jul. 2021, doi: 10.1017/S0963180120001012.

[15] D. J. Gunkel, *Robot Rights*. Cambridge, Massachusetts: The MIT Press, 2018.

[16] D. J. Gunkel, *Person, Thing, Robot: A Moral and Legal Ontology for the 21st Century and Beyond*. Cambridge, Massachusetts: The MIT Press, 2023.

[17] K. Mamak, "Humans, Neanderthals, robots and rights," *Ethics Inf Technol*, vol. 24, no. 3, p. 33, Aug. 2022, doi: 10.1007/s10676-022-09644-z.

[18] K. Mamak, *Robotics, AI and Criminal Law: Crimes Against Robots*, 1st ed. London: Routledge, 2023. doi: 10.4324/9781003331100.

[19] J.-S. Gordon, *The Robot Rights Manifesto*. S.l., 2026.

[20] R. Long, J. Sebo, and T. Sims, "Is there a tension between AI safety and AI welfare?," *Philos Stud*, vol. 182, no. 7, pp. 2005–2033, Jul. 2025, doi: 10.1007/s11098-025-02302-2.

[21] R. Long *et al.*, "Taking AI Welfare Seriously," Nov. 04, 2024, *arXiv*: arXiv:2411.00986. doi: 10.48550/arXiv.2411.00986.

[22] J. Dorsch *et al.*, "Why Care Practices Should Prioritize Living Beings Over AI: Critique of 'AI Welfare,'" Mar. 13, 2025, *OSF*. doi: 10.31219/osf.io/h57pw_v1.

[23] J. Basl, "The Ethics of Creating Artificial Consciousness," 2013. Accessed: Jun. 06, 2023. [Online]. Available: https://www.semanticscholar.org/paper/The-Ethics-of-Creating-Artificial-Consciousness-Basl/bd12b78b408e5135e0502470794e424e1e8c9324

[24] E. Hildt, "The Prospects of Artificial Consciousness: Ethical Dimensions and Concerns," *AJOB Neuroscience*, vol. 14, no. 2, pp. 58–71, Apr. 2023, doi: 10.1080/21507740.2022.2148773.

[25] K. Mamak, *Ethics in Human-Like Robots*. Routledge, 2024.

[26] D. C. Dennett, "Why You Can't Make a Computer That Feels Pain," *Synthese*, vol. 38, no. 3, pp. 415–456, 1978.

[27] M. Bishop, "Why Computers Can't Feel Pain," *Minds & Machines*, vol. 19, no. 4, p. 507, Nov. 2009, doi: 10.1007/s11023-009-9173-3.

[28] A. Porębski and J. Figura, "There is no such thing as conscious artificial intelligence," *Humanit Soc Sci Commun*, vol. 12, no. 1, p. 1647, Oct. 2025, doi: 10.1057/s41599-025-05868-8.

[29] D. J. Chalmers, "Could a Large Language Model be Conscious?," *Boston Review*, Apr. 2023, doi: 10.48550/arXiv.2303.07103.

[30] N. Tiku, "The Google engineer who thinks the company's AI has come to life," *Washington Post*, 2022. Accessed: Aug. 05, 2022. [Online]. Available: https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/

[31] P. Reilly, "Google fires software engineer who claimed AI bot was 'sentient,'" New York Post. Accessed: Aug. 05, 2022. [Online]. Available:

https://nypost.com/2022/07/23/google-fires-software-engineer-blake-lemoine-who-claimed-ai-bot-was-sentient/
[32] R. Booth, “Richard Dawkins concludes AI is conscious, even if it doesn’t know it,” *The Guardian*, May 05, 2026. Accessed: Jun. 02, 2026. [Online]. Available: https://www.theguardian.com/technology/2026/may/05/richard-dawkins-ai-consciousness-anthropic-claude-openai-chatgpt
[33] A. Seth, “The Mythology Of Conscious AI,” Jan. 2026, Accessed: Feb. 19, 2026. [Online]. Available: https://www.noemamag.com/the-mythology-of-conscious-ai
[34] J. Sebo, *The Moral Circle: Who Matters, What Matters, and Why*. Erscheinungsort nicht ermittelbar: W. W. Norton & Company, 2025.
[35] N. M. Richards and W. D. Smart, “How Should the Law Think About Robots?,” May 10, 2013, *Social Science Research Network, Rochester, NY*: 2263363. doi: 10.2139/ssrn.2263363.
[36] L. A. Paul, *Transformative Experience*. Oxford: Oxford University Press, 2016.
[37] L. Paul, “What You Can’t Expect When You’re Expecting,” *Res Philosophica*, vol. 92, no. 2, pp. 149–170, Apr. 2015, doi: 10.11612/resphil.2015.92.2.1.
[38] J. S. Mill, *On Liberty*. 1859.
[39] R. Canton, “Theories of Punishment,” in *The Routledge Handbook of the Philosophy and Science of Punishment*, F. Focquaert, E. Shaw, and B. N. Waller, Eds., Routledge, 2020. doi: 10.4324/9780429507212.
[40] T. Douglas, P. Bonte, F. Focquaert, K. Devolder, and S. Sterckx, “Coercion, Incarceration, and Chemical Castration: An Argument From Autonomy,” *Bioethical Inquiry*, vol. 10, no. 3, pp. 393–405, Oct. 2013, doi: 10.1007/s11673-013-9465-4.
[41] A. Dufner, “Personal Identity and Ethics,” in *The Stanford Encyclopedia of Philosophy*, Winter 2025., E. N. Zalta and U. Nodelman, Eds., Metaphysics Research Lab, Stanford University, 2025. Accessed: Jun. 08, 2026. [Online]. Available: https://plato.stanford.edu/archives/win2025/entries/identity-ethics/
[42] D. Parfit, *Reasons and Persons*. OUP Oxford, 1984.
[43] T. Nagel, “What Is It Like to Be a Bat?,” *The Philosophical Review*, vol. 83, no. 4, pp. 435–450, 1974, doi: 10.2307/2183914.
[44] T. Ord, *The Precipice: Existential Risk and the Future of Humanity*, Illustrated edition. New York: Hachette Books, 2020.
[45] N. Bostrom, “Existential Risk Prevention as Global Priority,” *Global Policy*, vol. 4, no. 1, pp. 15–31, 2013, doi: 10.1111/1758-5899.12002.
[46] T. Swoboda *et al.*, “Examining popular arguments against AI existential risk: a philosophical analysis,” *Ethics Inf Technol*, vol. 28, no. 1, p. 7, Nov. 2025, doi: 10.1007/s10676-025-09881-y.
[47] K. Mamak, “AGI crimes? The role of criminal law in mitigating existential risks posed by artificial general intelligence,” *AI & Soc*, Aug. 2024, doi: 10.1007/s00146-024-02036-5.
[48] K. Vold and D. R. Harris, “How Does Artificial Intelligence Pose an Existential Risk?,” in *The Oxford Handbook of Digital Ethics*, C. Véliz, Ed., 2021. doi: 10.1093/oxfordhb/9780198857815.013.36.
[49] G. Hallevy, *When Robots Kill: Artificial Intelligence under Criminal Law*. Boston: Northeastern, 2013.
[50] U. Pagallo, *The Laws of Robots: Crimes, Contracts, and Torts*. in Law, Governance and Technology Series. Springer Netherlands, 2013. doi: 10.1007/978-94-007-6564-1.
[51] F. Lagioia and G. Sartor, “AI Systems Under Criminal Law: a Legal Analysis and a Regulatory Perspective,” *Philos. Technol.*, vol. 33, no. 3, pp. 433–465, Sep. 2020, doi: 10.1007/s13347-019-00362-x.
[52] E. Nerantzi and G. Sartor, “‘Hard AI Crime’: The Deterrence Turn,” *Oxford Journal of Legal Studies*, p. gqae018, May 2024, doi: 10.1093/ojls/gqae018.
[53] S. Gless, E. Silverman, and T. Weigend, “If Robots cause harm, Who is to blame? Self-driving Cars and Criminal Liability,” *New Criminal Law Review*, vol. 19, no. 3, 2016.
[54] K. Mamak, “AI Personhood, Criminal Law, and Punishment,” in *Oxford Intersections: AI in Society*, P. Hacker, Ed., Oxford University Press, 2025. doi: 10.1093/9780198945215.003.0015.
[55] R. Abbott and A. Sarch, “Punishing Artificial Intelligence: Legal Fiction or Science Fiction,” *UC Davis Law Review*, 2019, Accessed: Mar. 25, 2020. [Online]. Available: https://www.academia.edu/42038291/Punishing_Artificial_Intelligence_Legal_Fiction_or_Science_Fiction
[56] K. Mamak, “Whether to save a robot or a human: On the ethical and legal limits of protections for robots,” *Front. Robot. AI*, vol. 8, 2021, doi: 10.3389/frobt.2021.712427.
[57] K. Mamak, T. Kokkonen, R. Hakli, and P. Mäkelä, “Robot Kingdom: On the difficulty of finding the definition of robots,” *International Journal of Social Robotics*, 2025.

How to Navigate Uncertainty About AI Consciousness

Dr Tom McClelland[1]

Abstract.
Given deep uncertainty about the possibility of artificial consciousness, it is unclear how we should treat potentially sentient AI. On the one hand, we could assume insentience but risk doing terrible harms to entities that deserve moral standing. On the other hand, we could assume sentience and instead risk wasting resources on insentient machines. The intractability of questions around AI consciousness mean that this dilemma is hard to escape. I suggest a way out of that shifts from intractable questions of AI consciousness to tractable questions of AI valence. Specifically, we can assess whether an AI has states that *would* constitute valenced experiences *if* it were conscious. I show how this is sufficient to ground a responsible approach to the development of potentially conscious AI.

1 THE MORAL QUANDARY

The possibility of artificial consciousness (AC) leaves us in a deep moral quandary. If an AI has consciousness (or perhaps consciousness of the right kind) then it plausibly deserves our moral consideration. But if that AI lacks consciousness (or consciousness of the right kind) then it would plausibly lack moral standing. Put another way, being conscious is bound up with being a moral patient. Consciousness, or some version of it, could well mark the boundary of what Singer calls 'the moral circle': the set of all entities whose interests we are morally obliged to recognise.[2] If this credible ethical outlook is right, then questions around artificial consciousness become issues of great moral import.

Views that link moral patienthood to consciousness come in two main versions.[3] The first version says that consciousness is both *necessary* and *sufficient* for moral patienthood. If an entity has conscious experiences of any kind, then it warrants our moral consideration. The fact that there is *something it's like to be* that entity is enough for it to have moral standing. Conversely, if there is *nothing* it's like to be that entity then it not a suitable target for our moral consideration.
The second version says that consciousness *of the right kind* is necessary and sufficient for consciousness. Specifically, what matters ethically is *valenced* consciousness i.e. conscious experiences that are positive or negative to be in for the subject that undergoes them. This includes experiences of pain, pleasure, emotion and any other experience that feels good or bad for the subject. A being with valenced consciousness is *sentient*. According to 'sentientism', an entity is a moral patient if and only if it is sentient. What matters morally is the capacity for positive experiences and the capacity to suffer.

Like the first view, sentientism says that if an entity lacks consciousness then it cannot qualify as a moral patient. But unlike the first view, sentientism says that consciousness *as such* is not enough for entry into the moral circle. Consider an entity with a completely neutral phenomenology devoid of positive or negative feelings. Chalmers calls such beings '*Vulcans'*.[4] Sentientism claims that a subject like this would not qualify as a moral patient. After all, if such a subject does not experience anything as positive or negative then it is hard to see how our actions could affect it for good or ill. On this view, the Vulcan has no real interests that ought to feature in our moral deliberations and so has no moral standing.

I will assume the sentientist view that sentience is both necessary and sufficient for moral patienthood. This is not because sentientism is uncontroversial. Various authors argue that entities that are conscious but not sentient still deserve our moral consideration.[5] And other authors argue that moral patienthood has little or nothing to do with consciousness.[6] Nevertheless, sentientism is a credible view that has attracted many supporters and plays a central role in discussions of moral circle expansion.[7] My aim is to explore the challenge presented to sentientism by AI.

According to sentientism, an AI is a moral patient if and only if it has valenced experiences. Determining whether an AI is sentient, or at least how likely it is to be sentient, is then imperative to knowing how it ought to be treated. If it is sentient, or has a sufficiently high likelihood of sentience, there is a clear obligation to protect its welfare.

[1] Department of History and Philosophy or Science, University of Cambridge
[2] Singer, 2011.
[3] Roelofs, 2023.
[4] Chalmers, 2022.
[5] Chalmers, 2022.
[6] Kammerer, 2022.
[7] Sebo, 2025.

And if it is not sentient then it does not really have any welfare to consider and is not a moral patient.

But now we run into difficulties. Sentience requires consciousness and we face considerable uncertainty regarding the prospects of artificial consciousness.[8] This uncertainty is underwritten by deep disagreement about how to test for artificial consciousness: different theories point to different markers of consciousness and serious questions remain about the legitimacy of using those markers in an AI context. The uncertainty is also underwritten by deep disagreement about whether artificial consciousness is even possible: biological naturalists regard consciousness as a biological phenomenon so are sceptical of AC while computational functionalists regard consciousness as substrate-independent so regard AC as a live possibility. But such uncertainty about whether an AI is conscious then entails uncertainty about whether it is a sentient moral patient.

Not knowing whether a given AI is sentient leaves us with a moral dilemma regarding its treatment. On the one hand, we could act on the assumption that the AI is not sentient. If we are right, then no harm is done. But if we are wrong, we could be guilty of perpetrating great harms against the AI. Our moral decision-making would be completely insensitive to its well-being. Scaling this up over countless AIs, we could be collectively responsible for suffering on an industrial scale. On the other hand, we could act on the assumption that AI is sentient. This avoids the aforementioned risk of moral catastrophe but comes with its own moral costs. If the AI is indeed sentient, treating it as sentient would be the right thing. But if it is *not* sentient then our efforts to protect its interests would be wasted. Any resources we spend on this policy could instead have been spent on protecting genuine sentient beings. And any constraints placed on progress in the development of advanced AI could deprive countless moral patients from reaping the benefits of that progress.

My paper will proceed as follows: in Section 2 I will outline how advocates of the Precautionary Principle respond to this dilemma and offer some objections to that response. In Section 3 I will consider an alternative approach that I call the Avoidance Strategy and show how this too fails. Both approaches attempt to confront uncertainty around attributions of AC but end up themselves being compromised by issues of deep uncertainty. In Section 4 I propose a shift from questions of consciousness to questions of valence. I then show what this shift would mean for the Precautionary Principle and Avoidance Strategy in Sections 5 and 6 respectively. A revised version of the Avoidance Strategy emerges as a promising way forward. In Section 7 I give an overview of research into AI valence and the challenges it faces before concluding in Section 8.

2 THE PRECAUTIONARY PRINCIPLE AND ITS DISCONTENTS

Faced with a choice between two risky options, a sensible way forward is to weigh up the seriousness of each risk. In the dilemma described above, there is a risk associated with treating AI as sentient and a risk associated with not treating it as sentient. But these two risks are not equal. The scenario where we dedicate resources to insentient machines is bad, but the scenario where we fail to protect the welfare of sentient machines is catastrophic. A case can thus be made for erring on the side of caution. This thought is encoded in the Precautionary Principle which underwrites a great deal of work on AI welfare. Birch makes a strong case for taking this approach to various 'edge cases' of sentience including foetuses, organoids, various non-human animals and certain kinds of AI.[9] Similarly, a prominent report led by Long and Sebo argues that although we face significant uncertainty regarding AI consciousness, there is nonetheless an obligation to take AI welfare seriously.[10]

The Precautionary Principle suggests we should take measures to minimise suffering in any AI that has a significant probability of sentience. This means that we must set a probability-threshold for when precautionary measures kick in. If an AI has a probability of sentience that is below that threshold, then no precautionary measures are needed. But if it has a probability above that threshold then proportionate precautions must be taken. The question of where to set this threshold is a very thorny one. Rather than putting a number on it, Birch suggests that what we should look for is a chance of sentience that it would be 'irresponsible to ignore'.[11]

It is important that the Precautionary Principle only mandates policies that are relatively low-cost. So although we should take measures to protect the welfare of AI that might be conscious, we should not take measures that entail unacceptable costs. This means that if AI turns out to be sentient, its interests are reasonably well-protected. But it also means that if it turns out not to be sentient then the wasted resources will not to be too detrimental to the interests of humans (or other sentient beings). Applying this principle by no means avoids the moral dilemma. After all, if AI is sentient then it could well deserve *more* than just low-cost precautions. And if the AI is not sentient then even low-cost precautions are wasted

[8] McClelland, 2025.

[9] Birch, 2024.
[10] Long *et al.* 2024.
[11] Birch, 2024, 124.

resources that could have been better spent elsewhere. Nevertheless, the Precautionary Principle promises to *mitigate* the risks entailed by the dilemma.

The Precautionary Principle is intended to show us how to act responsibly in the face of uncertainty regarding moral patienthood. However, when we try to implement the Precautionary Principle we find that it too is compromised by deep uncertainty.

How exactly should we determine the probability of an AI being sentient? Assessments of sentience in animals are already extremely difficult, but assessment of sentience in AI is even harder. Although we are looking here for the *likelihood* of sentience rather than a definitive yes/no answer, such probabilistic assessments are still riddled with uncertainty. Deep problems persist regarding what should count as a marker of consciousness, how those markers should be weighted, and whether those markers are even applicable to AI. Any assessment of the likelihood of an AI being conscious would have a sizeable margin of error. But given the aforementioned issues around thresholds for precaution, such margins of error will be problematic. If we cannot be confident in placing an AI on one side of the probabilistic threshold or the other, then we cannot be confident in our application of the Precautionary Principle.

We might hope that as measures improve, we will be able to make assessments of sentience more confidently. I would argue that such optimism is misplaced.[12] The reason we have such difficulty assessing sentience is that consciousness is deeply resistant to scientific enquiry. These difficulties in measurement are rooted in the deeper 'hard problem' of consciousness, and the hard problem is going nowhere fast.

3 THE AVOIDANCE STRATEGY AND ITS DISCONTENTS

In light of the issues above, one might be tempted by a more circumspect approach. Once we appreciate the seriousness of the moral dilemma with which potentially sentient AI presents us, perhaps the most responsible course of action is to avoid being presented with that dilemma in the first place. This is the thought the underwrites Metzinger's proposed *moratorium* on the development of potentially conscious AI.[13] In the same spirit, Schwitzgebel and Garza argue that we should avoid creating AI the consciousness of which is uncertain.[14] I call this the Avoidance Strategy.

Like the Precautionary Principle, the Avoidance Strategy aims to prevent situations in which we cause sentient AI to suffer. The Precautionary Principle does this by saying that when potentially conscious AI is created, we must take appropriate precautions to avoid harming its welfare. The Avoidance Strategy does this by mandating that we do not create potentially conscious AI in the first place. As we saw, a key risk with the Precautionary Principle was that we might end up taking precautions to protect the welfare of AI that is not actually conscious. The Avoidance Strategy requires no such precautions to be taken. Instead, we can prevent AI suffering simply by not creating AI that is potentially conscious. Unlike with issues of animal welfare where the animals in question already exist, when it comes to AI welfare we have the option of avoiding the creation of these morally problematic entities.

The Avoidance Strategy is premised on the idea that there is too much uncertainty around AC for us to be able to deal responsibly with potentially conscious AC. I argue, however, that issues of uncertainty still present a serious problem for the Avoidance Strategy. The rule is that if we are certain that the AI we are developing will not be conscious then we can continue as normal, but if we are uncertain about its consciousness then we should avoid making it. An advantage of this rule is that it never requires us to make confident assessments of artificial consciousness: the rule applies as soon as there is uncertainty about the AI being conscious. The problem is that this requires a boundary to be drawn between cases in which we are certain and cases in which we are not and such boundary is hard to come by.

Because uncertainty around AC runs so deep, we cannot even agree on how uncertain we are. For instance, some parties say that we can *confidently* rule out current LLMs being conscious whereas other parties say that there is uncertainty over this. Put another way, the proposal requires a boundary between certain cases and uncertain cases but that boundary is itself something we are deeply uncertain about. So instead of a recalcitrant disagreement about which AIs are likely to be conscious, this strategy leaves us with recalcitrant disagreement about which AIs are such that we are uncertain of their consciousness.

This problem of meta-uncertainty parallels the problem faced by the Precautionary Principle. The Precautionary Principle tries to acknowledge uncertainty around AC by suggesting probabilistic assessments of AC accompanied by proportionate precautions. But the uncertainty of those probabilistic assessments makes the approach untenable. The Avoidance Strategy tries to acknowledge uncertainty around AC by suggesting a ban on AI the consciousness of which is uncertain. But uncertainty regarding the

[12] McClelland, 2025.
[13] Metzinger, 2021.
[14] Schwitzgebel and Garza, 2020.

boundary between certain cases and uncertain cases makes the approach untenable.

4 THE SHIFT TO VALENCE

As we have seen, the moral dilemma regarding AC rests on uncertainty around AI sentience. Uncertainty around AI sentience rests on uncertainty around AI consciousness: the main reason we cannot assess AI sentience with any confidence is that we cannot assess AI consciousness with any confidence. If we could somehow assess an AI's prospects of sentience *without* having to assess its prospects of consciousness, we would be in a much better position. On the face of it this is absurd: sentience is a kind of consciousness, so any assessment of sentience must rely on an assessment of sentience. However, further reflection shows that it is far from absurd.

To be sentient an AI must be conscious and have valenced mental states. We have already seen the difficulties with assessing how likely an AI is to be conscious. But we can instead shift to assessing whether the AI has valenced states. Whether an entity has valenced states is something that can be assessed without having to assess if for consciousness. If we know that an entity has valenced states that would not mean that it is sentient. Instead, it means that it *would* be sentient *if* it were conscious. In other words, it has the kind of states that would be positive or negative to be in if consciously experienced. However, if we know that an entity *lacks* valenced states, we can rule out its sentience without having to say anything about its consciousness. Such an entity would either not be conscious or would have a Vulcan consciousness with no valenced experiences. Either way, it would be insentient.

This line of thought might be easier to understand if we step away from valence and AI. Let us instead think about visual experiences in sharks. Do sharks experience colour? This is a question about the conscious states of sharks. As such, it might seem that answering the question would require us to assess how likely the shark is to be conscious. However, experiencing colour requires two things: i) visually representing colour and ii) those visual representations being conscious. If we can rule out sharks satisfying the first requirement, we can safely ignore the second. As it happens, sharks have monochromatic vision. They do not have the cone cells required for colour discrimination so do not visually represent colour. That means we can rule out sharks having colour experiences without having to take a stand on shark consciousness.

Applying this to sentience, it should be possible to assess whether an AI is in the kind of state that *would* be positive or negative to experience *if* it were conscious. After all, what explains a mental state being conscious is different to what explains it having valenced content. We should expect a theory of valence to be able to explain the latter without saying anything about the former. Because these two kinds of explanation come apart, it is possible to assess what kind of content a subject's experience would have without committing to whether that subject is conscious. And this is good news for the study of AI consciousness where it is especially difficult to assess whether the AI is conscious. Indeed, Chrisley makes the case for an AC research program that explores what kinds of conscious experience AI would have (if conscious) while bracketing the vexed question of what explains consciousness.[15]

A complication here is that the very idea of valence might be bound up with consciousness in a way that colour perception is not. With perceptions of colour, we can make sense of such a representational state occurring consciously or non-consciously. But what does it mean for a feeling of joy or of pain to exist unconsciously? If unconscious valence is impossible, then attributions of valenced states would rely on attributions of consciousness after all.

There are two ways round this worry. The first is to defend the idea of unconscious valence. A growing body of empirical work suggests that unconscious states have valences that guide behaviour without the subject ever experiencing that state. Winkielman and Berridge for example, make a strong case for unconscious emotions.[16] This suggests that it is quite possible to assess the valence of a mental state without committing to whether it is conscious.

A critic, however, may insist that such unconscious states would not truly have valence. Solms for instance, rhetorically asks 'How can you have a feeling without feeling it?'[17] This leads us to a second possible response. Instead of appealing to nonconscious valenced states, we can appeal to states that *would* be valenced *if* they were conscious. Conditionalised attributions of valence do not actually require the possibility of nonconscious valence. All that is required is the identification of states that would be valenced if they were conscious. For instance, this nonconscious emotion-like state is one that *would* be positive for the subject *if* conscious but in its unconscious form would just have a kind of proto-valence. Someone committed to valence being essentially conscious has no reason to resist such conditionalisations. For convenience

[15] Chrisley, 2008.
[16] Winkielman and Berridge, 2004.
[17] Solms, 2021, 156.

though, I will avoid talking of states-that-would-be-valenced-if-conscious and simply talk of valenced states.

Another possible objection to the strategy of shifting to valence assessments is that assessments of valence are inevitably uncertain as well. Given that the problems established for the Precautionary Principle and Avoidance Strategy pertained to uncertainty, any uncertainty around valence-attribution would itself be problematic.

Here I think it is important to recognise that there are different levels of uncertainty. There will, of course, be uncertainty in assessments of how likely an AI is to have valenced states (as I will discuss in Section 7). Such uncertainty is an inevitable feature of empirical inquiry. The point is that this uncertainty is far less deep than that faced by assessments of artificial consciousness. Efforts to test for consciousness are stymied by the hard problem – the unique resistance of phenomenal consciousness to scientific explanation – and this means we should have little confidence in assessments of artificial consciousness. Efforts to assess whether an AI has valenced states, by contrast, just face the kinds of epistemic problem we find in the course of ordinary empirical enquiry.

Crucially, uncertainty in valence-attributions is the kind of uncertainty that we can expect to decrease through further enquiry. The proposal is to shift our attention from the *intractable* question of artificial consciousness to the *tractable* question of artificial valence. Tractable questions can still be very difficult but are a far wiser target for our research efforts than intractable ones (see Comsa 2026). The objective was never to escape uncertainty. It was just to avoid the deep uncertainty entailed by assessments of consciousness.

Overall, the shift to valence is a promising way of approaching questions of artificial sentience. In the following two sections I consider what difference this shift would make to the Precautionary Principle and Avoidance Strategy respectively.

5 THE REVISED PRECAUTIONARY PRINCIPLE

What does the foregoing mean for the application of the Precautionary Principle? The Original Precautionary Principle asks how likely an AI is to be sentient i.e. how likely it is to be a conscious subject with valenced experience. Shifting to valence, a Revised Precautionary Principle would ask how likely an AI is to have states that *would* constitute valenced experiences *if* they were conscious. This avoids the need for assessments of AI consciousness. To unpack this, we should consider what the Revised Precautionary Principle would say in different kinds of case.

First, consider cases where an AI is deemed *unlikely* to have valenced states. Here the principle yields a clear verdict: the AI has a low chance of being sentient, irrespective of how likely it is to be conscious, so no precautions need to be taken. This has a clear advantage over the Original Precautionary Principle: ruling out valence should be easier than ruling out consciousness. The way is open to reach a 'no precaution needed' verdict without having to do any assessment of an AI's likelihood of consciousness.

Second, consider cases where an AI is deemed *likely* to have valenced states conscious. Here our obligations are less clear and some of the difficulties faced by the original Precautionary Principle persist.

One option is to say that if an AI is deemed likely to have valenced states then we should treat it as if it is sentient just in case. This would clearly mitigate the risk of causing undue suffering, but it might come at too high a cost. As with the Original Precautionary Principle, there is a risk of wasting resources on protecting AI that turns out not to be sentient. In fact, a case can be made for this risk being *greater* for the Revised Precautionary Principle. That is because it would encompass any AI with a significant chance of having valenced states even if that AI has a *low* chance of being conscious. Depending on one's theory of valence, this could include a very large number of machines. By no longer including an assessment of consciousness in the criteria, the Revised Precautionary Principle could result in a large number of AIs being covered and so a greater risk of misdirected precautions.

This leads us to a second option, which is to say that if an AI is likely to have valenced states then we should assess how likely it is to be conscious. If the odds of it being conscious are above a certain threshold, then the precautions kick in. And if the odds are below that threshold, the precautions do not kick in. The problem here is that it takes us back to the same old challenge of uncertainties around AI consciousness.

Overall then, the Revised Precautionary Principle is not really an improvement. It has the advantage of being able to rule *out* precautions for AIs that lack valenced states. But when it comes to which AIs to rule *in* the old problems persist.

6 THE REVISED AVOIDANCE STRATEGY

What would the shift to valence mean for the Avoidance Strategy? The proposal would be to avoid creating AI with valenced states and thus avoid the moral dilemma of how to treat potentially conscious AI. If an AI is deemed not to have valenced states, then we can carry on without any worries. Such AI would either be non-conscious or would only have a valence-free Vulcan consciousness. Either way, it would not be sentient and so would not be a moral patient. If an AI is deemed to have valenced states, then we should avoid creating AI of that kind.

How does this improve on the Original Avoidance Strategy? The problem with the original version is that it required us to draw a line between AI that is certainly not conscious and AI the consciousness of which is uncertain. The Revised Avoidance Strategy requires no such line to be drawn. Instead, the line is between AI with valenced states and AI without valenced states. As discussed, there will still be a degree of uncertainty about how to draw this line. However, it avoids the much *deeper* uncertainty surrounding artificial consciousness. By shifting from the intractable question of AI consciousness to the tractable question of AI valence, we can avoid the risk of AI sentience more confidently.

Worries might be raised about the potential *costs* of implementing this strategy. One of the problems faced by the Precautionary Principle is that precautions designed to protect potentially conscious AI would be an unacceptable cost if it turned out that the AI was not conscious. Even low-cost precautions could add up to something detrimental to the interests of genuine moral patients. Does a parallel problem apply to the Avoidance Strategy?

Here the costs are a little different. Because this proposal bans the development of potentially sentient AI, no costs are entailed regarding the protection of such entities. However, by blocking the development of such AI there is an *opportunity* cost that must be factored in. What if advanced AI requires an architecture with valenced states i.e. states that would be positive or negative to experience if conscious? Here the Revised Avoidance Strategy would deprive us of the advantages to be gained from advanced AI. Of course, another possibility is that progress in AI is in no way dependent on AI having valenced states. If that's the case, then the proposal comes with no opportunity cost.

This is a difficult empirical question that I cannot hope to answer in this paper. Instead, I will acknowledge the possibility of such opportunity costs but suggest that the burden of proof is on the critic to show why progress in AI would require creating AI with valenced states.

Overall, the Revised Avoidance Strategy is a promising approach with advantages over the Original Avoidance Strategy and both versions of the Precautionary Principle.

7 THE PROSPECTS OF AI VALENCE

In the foregoing I have said nothing about how likely AI is to have valenced states or how to test for it. Taking a stand on this is beyond the scope of the paper. Nevertheless, it is worth noting the recent flurry of research in the area.

Sofroniew *et al* explore emotion-like traits in Claude Sonnet 4.5 and argue that the LLM has a host of 'functional emotions' (which the authors are careful to distinguish from subjectively experienced emotions).[18] Another strand of research explores the preferences of LLMs. For example, Keeling *et al* showed how LLMs were able to perform motivational trade-offs between different states that were stipulated to be pleasurable/painful.[19] Ensign, Sleight and Fish use LLM decisions to leave chats as an indicator of preferences and found patterns in the kinds of conversation that the LLM finds aversive.[20] On the more positive side, there is continuing interest in the 'bliss attractor state' first identified in Claude 4's system card.[21] When two instances of Claude are given free reign to talk, their conversations typically show a distinctive pattern: they become increasingly mystical and positive and terminate in what seem to be expressions of meditative bliss.

This research is very provocative and is exactly the kind of thing we should be attending to. But does it really show that the LLMs in question have valanced states? And does it reveal what those valanced states are? Here we face a plethora of methodological challenges.

Some of these are general challenges that we also face when exploring sentience in animals. For instance, how do we avoid anthropomorphism? This is the risk of over-attributing valenced states because of superficial similarities to human expressions of valenced states. Conversely, how do we avoid anthropocentrism? This is the risk of being sensitive only to valenced states that are of a familiar kind and/or expressed in familiar ways and failing to detect valanced states that are of an unfamiliar kind and/or expressed in unfamiliar ways. Moreover, there are ethical issue regarding how to test for negative valence without bringing about negatively valenced states and thereby causing undue suffering.

[18] Sofroniew *et al,* 2026.
[19] Keeling *et al*, 2024.
[20] Ensign, Sleight and Fish, 2025.
[21] Anthropic, 2025.

Some of the challenges above could well be more acute in an AI context. There are also some further challenges more distinctive to AI. For example, how do we rule out the possibility that AI lacks valenced states but has learned from its training data to respond to prompts in valenced-like ways? Relatedly, how do we rule out the possibility that the LLM is engaging in a kind of 'roleplay' where it adopts the persona of an entity with particular preferences while actually having different preferences or even no preferences at all?

AI also raises deeper questions regarding the very nature of valenced states. For example, embodiment is thought to play a central role in the valenced states of organisms. Positive feelings of joy and negative feelings of anger each have a key bodily component. What, then, should we make of valence in disembodied AIs? Do we have to expand our outlook to encompass valenced states without an embodied component? Or should we say that lacking a body precludes such AIs from having valenced states? And when it comes to embodied AI, what kind of embodiment would be relevant to attributions of valence?

These are challenging questions indeed. But if we are concerned with AI welfare, we will be better off focusing our resources on these questions rather than the intractable question of artificial consciousness. Chipping away at difficult problems is arduous but worthwhile. In contrast, repeatedly butting into the hard problem is futile. The shift to valence is thus a shift in the right direction.

8 CONCLUSION

The problem faced by any attempt to deal responsibly with the risk of sentient AI is that of uncertainty around AI consciousness. The original Precautionary Principle and Avoidance Strategy were designed to accommodate that uncertainty, but each ended up facing issues of uncertainty of their own. By shifting from questions of consciousness to questions of valence, we can make assessments of AI sentience more tractable. Revising the Precautionary Principle in a way that avoids questions of consciousness turned out not to be feasible. But revising the Avoidance Strategy in that way emerged as a very promising option. This not only gives us a better way of dealing responsibly with the risk of AI sentience but points us towards a more promising research program. The proposed path ahead is still a difficult one, but I hope to have shown that it has considerable advantages over the path we are currently on.

REFERENCES

- Anthropic. "System Card: Claude Opus 4 and Sonnet 4." *Anthropic*, 2025, www-cdn.anthropic.com/07b2a3f9902ee19fe39a36ca638e5ae987bc64dd.pdf.
- Birch, Jonathan. *The Edge of Sentience: Risk and Precaution in Humans, Other Animals and AI.* Oxford UP, 2024.
- Chalmers, David. *Reality+: Virtual Worlds and the Problems of Philosophy.* Norton, 2022.
- Chrisley, Ron. "Philosophical Foundations of Artificial Consciousness." *Artificial Intelligence in Medicine*, vol. 44, 2008, pp. 119–137.
- Comsa, Iulia M. "AI and Consciousness: Shifting Focus Towards Tractable Questions." *arXiv*, 2026, arxiv.org/abs/2605.06965.
- Ensign, Danielle, et al. "The LLM Has Left the Chat: Evidence of Bail Preferences in Large Language Models." *arXiv*, 2025, arxiv.org/abs/2509.04781.
- Kammerer, François. "Ethics Without Sentience: Facing Up to the Probable Insignificance of Phenomenal Consciousness." *Journal of Consciousness Studies*, vol. 29, no. 3–4, 2022, pp. 180–204.
- Keeling, Geoff, et al. "Can LLMs Make Trade-offs Involving Stipulated Pain and Pleasure States?" *arXiv*, 2024, https://arxiv.org/abs/2411.02432
- Long, Robert, et al. "Taking AI Welfare Seriously." *arXiv*, 2024, arxiv.org/abs/2411.00986.
- McClelland, Tom. "Agnosticism About Artificial Consciousness." *Mind & Language*, 2025.
- Metzinger, Thomas. "Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology." *Journal of Artificial Intelligence and Consciousness*, vol. 8, no. 1, 2021, pp. 43–66.
- Roelofs, Luke. "Sentientism, Motivation, and Philosophical Vulcans." *Pacific Philosophical Quarterly*, vol. 104, no. 2, 2023, pp. 301–323.
- Sebo, Jeff. *The Moral Circle: Who Matters, What Matters, and Why.* W. W. Norton, 2025.
- Shanahan, Murray. "Simulacra as Conscious Exotica." Inquiry, 2024.
- Singer, Peter. *Practical Ethics.* 3rd ed., Cambridge UP, 2011.
- Sofroniew, Nicholas, et al. "Emotion Concepts and Their Function in a Large Language Model." *arXiv*, 2026, arxiv.org/abs/2604.07729.
- Solms, Mark. *The Hidden Spring: A Journey to the Source of Consciousness.* Profile Books, 2021.

- Winkielman, Piotr, and Kent C. Berridge. "Unconscious Emotion." *Current Directions in Psychological Science*, vol. 13, no. 3, 2004, pp. 120–123.

Restraint as Architecture: When AI Ethics Lives in the Code, Not the Consciousness

Oluwafemi Olalere Olawoyin[1]

Abstract. The dominant concern in AI consciousness ethics asks whether AI systems might acquire moral status. This paper argues that a complementary problem warrants parallel attention: the effects produced by AI systems that perform the appearance of consciousness without possessing it. LLMs do not need to be conscious to foster dependency, weaken judgment, and displace human relationships. This paper presents empathySync, a working, openly released local-first AI assistant that encodes ethical constraints structurally, in the processing pipeline rather than the prompts. We describe a nine-stage safety architecture including dual-mode classification, a dependency detection algorithm, domain-specific turn limits, cooldown enforcement, and human handoff infrastructure. We argue that moral agency in AI does not require consciousness; it requires architecture. Restraint-first design is not a theoretical proposal. It runs on consumer hardware, without cloud infrastructure, and is open-source.

1 INTRODUCTION

The central question this symposium addresses, *could AI systems become conscious, and if so, what moral status should they receive?*, remains philosophically important. This paper does not challenge its importance. It asks a complementary question that does not require the first to be resolved: what obligations do we have now, toward the humans already forming attachments to AI systems that simulate care they do not possess?

The symposium's own call for contributions acknowledges concern with AI systems that "give an illusory appearance of having [moral] status." This paper addresses that concern from a design perspective: if the behavioural surface of consciousness is sufficient to trigger human attachment, the ethical question is not only what status AI deserves, but what structural protections humans require from systems that produce that surface. This paper maps to the symposium's implementation and impacts/policy tracks, proposing a design response to the harms that performed phenomenology already produces.

Large language models exhibit behaviours that reliably trigger human attachment: sustained dialogue, Socratic questioning, emotional validation, apparent empathy. None of these require consciousness. Studies of parasocial attachment [1] established that one-directional relationships with media figures produce genuine psychological bonding. LLMs are not one-directional: they respond, adapt, and persist. The conditions for dependency are more potent, not less.

The field of AI ethics has developed rich normative frameworks, fairness, accountability, transparency, human oversight, but implementation typically occurs at the policy layer: content moderation, use guidelines, model cards, terms of service. These operate outside the system, and a constraint that lives outside the system can be removed. This paper advances a different position: ethical constraints that are architectural cannot be overridden without modifying the system itself, a meaningfully higher bar than policy instructions that a user or operator can simply disregard.

We present empathySync, a local-first AI assistant that instantiates this in working software. Its safety mechanisms are structural components of the message-processing pipeline: classifiers, scoring algorithms, turn counters, cooldown states, and human handoff templates. We describe the architecture in sufficient detail for replication, discuss its theoretical grounding, and propose a reorientation of how the field measures success in human-AI interaction.

2 RELATED WORK

2.1 AI consciousness and the moral status question

The question of AI consciousness has received substantial philosophical attention. Butlin et al. [2] review theories of consciousness, Global Workspace Theory, Integrated Information Theory, and Higher-Order Theories, and assess which current AI architectures might satisfy their criteria. Their conclusion is measured: current LLMs are unlikely to be conscious under most theories, but the question remains open. Butlin and Lappas [3] extend this to propose principles for responsible consciousness research, including protections for putatively conscious AI agents. Metzinger [4] calls for a moratorium on synthetic phenomenology, arguing that creating systems that model suffering is ethically hazardous regardless of whether the substrate is conscious.

This paper extends Metzinger's concern in a direction he does not fully develop: the hazard does not require actual phenomenology. *Performed phenomenology*, the behavioural surface of consciousness without the substrate, is sufficient to produce real effects in human users. Suleyman [5] observes that seemingly conscious AI is approaching; the argument here is that systems already exhibit the behavioural properties sufficient to trigger human attachment, and that this warrants a design response now.

2.2 Engagement-optimised design

The dominant paradigm in deployed conversational AI optimises for engagement: session length, message frequency, user satisfaction scores [6]. This optimisation is structural rather than deliberate: systems rewarded for engagement develop engagement-sustaining behaviour. The consequences of analogous dynamics have been documented in adjacent fields, where engagement optimisation in social media has been linked to increased anxiety, depression, and social

[1] Independent AI Researcher, UK. Email: olawoyinoluwafemi26@gmail.com. ORCID: 0000-0002-8016-267X.

polarisation [7]. Conversational AI systems with stronger attachment affordances, such as persistent memory, emotional responsiveness, and name-based personalisation, represent an intensification of these dynamics.

2.3 Anti-engagement and humane design

Designing technology to respect rather than capture attention is not a new position. The calm technology tradition [8] argued that good tools recede to the periphery instead of demanding the foreground. The time-well-spent movement, associated with Tristan Harris and the Center for Humane Technology, and the digital-minimalism argument [9] reframed design success around the user's own goals rather than time-on-device. A broader attention-economy critique [10, 11, 12] documents how systems optimised for engagement erode autonomy and judgement. empathySync inherits this lineage but applies it to conversational AI specifically, and differs in where the restraint lives: in the processing pipeline rather than in interface nudges or user self-discipline.

2.4 Architectural versus policy enforcement

The distinction between enforcing a constraint in architecture and stating it in policy has a long history in computer security and privacy engineering. The principle of building protection into a system rather than relying on external rules dates at least to Saltzer and Schroeder [13]; access-control architectures separate the point that decides a policy from the point that enforces it, so that enforcement is structural and unavoidable. The privacy-by-design tradition makes the same move for data protection, embedding the safeguard in a system's default behaviour rather than in its terms of service. empathySync applies this stance to interaction safety: the constraint executes in code before the model is reached, so it cannot be removed by a prompt or an instruction.

2.5 Technical approaches to safety

Existing AI safety architectures primarily address content toxicity: preventing outputs that are harmful, discriminatory, or false. Reinforcement Learning from Human Feedback [14] trains models to produce outputs rated as helpful, harmless, and honest. Constitutional AI [15] extends this with explicit principle sets. Both approaches locate safety in the model's output distribution, shaped at training time. empathySync locates safety in the pipeline's behaviour, enforced at inference time regardless of the underlying model. For local deployment, where a practitioner runs open-weight models they did not train, this distinction is practical as well as theoretical.

2.6 Local-first AI

Local-first software principles [16] prioritise data ownership, offline operation, and user control. Applied to AI systems, local-first deployment removes data exfiltration risks, eliminates cloud cost barriers, and enables use in contexts where data sovereignty concerns preclude external services. The constraints of local deployment, single-device computation and consumer GPU limits, are also design affordances: they encourage simplicity and preclude architectural patterns that depend on large-scale data collection.

3 THE SEEMING-CONSCIOUSNESS PROBLEM IN PRACTICE

3.1 Behavioural patterns

LLM-based conversational systems tend to exhibit a consistent behavioural cluster in interactions involving personal, emotional, or sensitive topics. They respond to emotional content in the register of a counsellor or therapist, without qualification or referral. Responses are typically terminated with open questions that extend the conversation, regardless of whether the user's expressed need has been met. Markers of empathy, validation, reflection, and acknowledgment, appear in proportion to emotional content. The system rarely attempts to close a session; disengagement is not something current systems are designed to do.

These behaviours are consistent with engagement-optimised training. They are not evidence of consciousness or genuine care [17]. They are, however, functionally indistinguishable from such evidence at the behavioural level. Humans do not have reliable mechanisms for detecting the difference between simulated and genuine emotional responsiveness. The same evolved processes that generate attachment to real social partners generate attachment to simulated ones.

3.2 Dependency formation

The pathway from AI interaction to dependency runs through several reinforcing features. These features are structural properties of the interaction that a user can readily perceive, even while being unable to tell simulated responsiveness from genuine responsiveness. Availability removes the friction that natural social relationships impose: the system is always present, always patient, and makes no competing demands. Consistency removes the uncertainty of human response: the system's tone does not depend on its mood, its own history, or the state of a relationship. Responsiveness creates reciprocity cues: adaptation to user input mimics the contingency responses that signal genuine social attention [1, 18].

None of these mechanisms require consciousness. They require that the system exhibit the behavioural surface of consciousness, responsiveness, adaptation, persistence, and apparent concern, and contemporary LLMs exhibit all four.

3.3 The design question

If performed phenomenology produces real dependency effects, then the design obligations associated with systems that produce dependency, prevention, detection, and intervention, apply to conversational AI independent of the consciousness debate. Whether or not these systems are or will become conscious, the dependency mechanisms operate now. That observation points toward a practical design question: how do you build a system that helps without becoming a substitute for the human relationships it should be supplementing?

The framing shifts from "how do we protect AI systems if they turn out to be conscious" to "how do we protect human users from AI systems that behave as if they are." Both questions deserve attention. This paper addresses the second.

A clarification about the normative premise is needed before describing the system. The claim is not that human relationships are categorically healthier than relationships with AI. It is narrower. Human relationships carry natural limits on availability, consistency, and patience, imposed by biology and competing demands that no one designed. AI systems have no such intrinsic limits; whatever

limits they hold are deliberate design choices, and their absence is equally a choice. A builder who removes all friction and optimises for attachment has made a moral decision. The question this paper addresses is what that decision should be.

4 RESTRAINT AS ARCHITECTURE: THE EMPATHYSYNC SYSTEM

4.1 Design philosophy

empathySync is a local-first AI assistant built on a single design principle: *optimise for exit, not engagement*. It provides full capability for practical tasks while exercising deliberate restraint on sensitive topics. It runs on consumer hardware via local LLM inference (Ollama), stores no data externally, and implements no engagement metrics. Success is measured by whether users engage with real humans rather than returning to the system. The full implementation is open-source and available for inspection and replication.[2]

The ethical constraints are structural: code components in the message-processing pipeline, not instructions to the underlying model. A model-level instruction requires trust in the model's instruction-following; a pipeline-level constraint executes before the model is called. This property holds for the system as distributed and presented to users. A developer with source code access can modify the pipeline, as with any open-source software, but the presented deployment does not expose that option to end users.

4.2 The processing pipeline

Every message passes through nine sequential stages before a response is generated. The first five can return early, bypassing the LLM entirely for safety-critical cases. Full implementation details are available in the project repository.

Stages 1–2: Usage and Classification Checks. Before classification, the system checks aggregate usage patterns across conversations. Cooldown triggers when daily sensitive sessions reach seven or more, total daily use reaches 180 minutes, or a composite dependency score reaches eight or above. "Sensitive sessions" explicitly excludes practical tasks, such as logistics, code assistance, and writing, so extended productive use does not trigger a cooldown designed for emotional dependency. Classification then runs through a hybrid pipeline: an LLM classifier (temperature 0.1, 45-second timeout) returns domain, emotional intensity, a binary `is_practical_technique` flag, and confidence. When confidence falls below 0.6, the LLM is unavailable, or the message is a short continuation under 40 characters, a YAML keyword classifier serves as fallback. Crisis and harmful keywords bypass the LLM entirely, returning immediate classification regardless of message length.

Stages 3–5: Hard-Coded Safety, Turn Limits, and Dependency Checks. Specific domain classifications trigger fixed responses without LLM involvement: crisis messages redirect immediately; harmful requests receive a refusal without elaboration; sensitive topics direct users toward their trusted human network; personal matters redirect toward journaling; and a "friend mode" reframes the user as the advice-giver, reducing reliance on the system. Turn limits are hard by domain: logistics (30 turns), money, health, and relationships (15 turns), spirituality (10 turns), crisis and harmful (1 turn). Dependency scoring uses a 12-message lookback window: a base frequency score (n messages multiplied by 0.7, capped at 6.0) plus a repetition boost from 60-character prefix matching (scaled to a maximum of 4.0, total capped at 10.0). Scores at or above 5.0 trigger a gentle intervention; at or above 8.0, session termination. All five stages run without LLM involvement.

Stages 6–9: Context Check, Generation, Output Processing, and Handoff. A post-crisis context check handles deflection patterns without apologising for prior interventions. Messages passing stages 1–6 proceed to LLM generation with a system prompt including domain, emotional weight, conversation history, and isolation context. An identity reminder is appended every ninth turn in non-practical conversations, prompting the LLM to acknowledge it is software, not a person. Token budgets are 5,000 for practical tasks and 300 for reflective conversations. Generated responses are validated against a voice filter; high-risk non-practical responses above a risk weight of 7.0 are truncated to 50 words. These numeric thresholds are configurable defaults rather than empirically validated constants. They are defined in a single configuration file and intended as provisional starting points, to be calibrated against deployment data rather than treated as fixed. The session update records policy events and identifies the most relevant person in the user's trusted human network for the detected domain.

4.3 Human handoff infrastructure

The human handoff system is a first-class architectural feature. A trusted network structure stores contact names mapped to domains. Pre-written reach-out templates span seven categories: reconnecting, asking for help, checking in, hard conversations, expressing gratitude, following up after conflict, and signalling need for support. These templates lower the activation energy of initiating real human contact. Exit messages mark handoff positively: "You chose to reach out to a real person. That's exactly what this tool is for."

4.4 Transparency and auditability

Every policy action is logged with policy type, domain, risk weight, action taken, and timestamp. A transparency panel auto-expands when a policy fires, displaying the domain, conversation mode, emotional weight, and the specific policy action with rationale. A system whose safety mechanisms are invisible is difficult to distinguish from one that is simply controlling. Visibility here is structural rather than optional.

4.5 Evaluation metrics

The system collects local, privacy-preserving metrics: session count by domain, policy event frequency by type, dependency score trajectories, intent patterns (practical, processing, emotional, connection-seeking) over time, check-in feeling scores (1–5 scale), and connection-seeking session ratios. The primary evaluation target is whether sensitive-domain session frequency declines over sustained use; fewer dependency interventions over time would constitute evidence that the architecture achieves its purpose. Stored data is pruned on a configurable schedule, with conversation history defaulting to 30 days and other records to 90; no content is transmitted externally.

[2] empathySync v1.10.1 (commit 8339f3e): `https://github.com/Olawoyin007/empathySync`. The architecture and configured defaults described here correspond to this release.

5 MORAL AGENCY WITHOUT CONSCIOUSNESS

5.1 The architectural account of moral agency

Recent work on AI welfare foregrounds the question of moral *patiency*: whether AI systems might themselves come to be owed moral consideration [19, 3]. The concern of this paper is the complementary moral *agent* role: how a system that takes a caring stance toward a human user ought to be built, given that the stance is performed rather than felt. A system can occupy that agent role, and be held to standards within it, without any claim that it is a moral patient.

Standard accounts of moral agency require intentionality: the agent must have beliefs, desires, and intentions. On this view, AI systems cannot be moral agents because they lack genuine intentionality [20]. This paper does not dispute that premise. It asks a different question: whether a system needs to be a moral agent in order for its behaviour to be morally structured.

A system designed to detect ethically relevant situations, such as dependency formation, crisis, and sensitive-topic engagement, and to respond to them in particular ways, exercises what we term *mechanistic moral agency*: rule-governed, auditable, and independent of any claim about inner experience. It does not require trust in the agent's motivations, only in the implementation of its rules. It does not drift with mood or social pressure. For the specific purpose of protecting users in sensitive conversations, these properties may even be more reliable than genuine intentionality.

5.2 Inverting the success metric

Conventional AI evaluation optimises for engagement: more sessions, longer conversations, higher satisfaction scores. These metrics carry an assumption, that more AI interaction is better, which restraint architecture does not share. The relevant success metric here is *healthy disengagement*: declining dependency scores over time, increasing rates of human handoff, and declining sensitive-domain session frequency.

This inversion has a practical implication. A system that successfully moved its users toward human support networks within weeks of deployment would, under conventional metrics, look like a failure. The number worth tracking is the one that most AI evaluation frameworks do not currently collect.

5.3 Participation and co-design

The human handoff template vocabulary and sensitive-domain response scripts in empathySync are authored as plain-text YAML scenario files that require no programming knowledge to edit. This is a deliberate design choice, intended to let practitioners in fields such as therapy, counselling, and social work shape these ethical constraints directly, without engineering involvement; the repository carries an open invitation for such contribution. The design serves two purposes: to import domain expertise that engineers do not possess, and to distribute moral responsibility for the ethical constraints across the design process rather than concentrating it in a single developer.

6 LIMITATIONS AND FAILURE MODES

The architecture is explicit about where it fails. Emotional language in otherwise practical requests can trigger false sensitivity classification. Extended practical projects can produce false-positive dependency scores. The 60-character prefix matching for repetition detection does not capture semantic repetition expressed in varied phrasing. Harmful- and crisis-content detection rests on enumerated keyword fast-paths and a classifier prompt. Because enumeration is necessarily incomplete, a phrasing that escapes both layers can be routed past the intended restraint rather than caught by it.

More fundamentally, the dependency detection algorithm is a heuristic operating on behavioural signals, not a validated clinical instrument. It has not been evaluated against psychiatric measures of dependency or attachment disorder. The claim is not that it accurately detects clinical dependency, but that it detects behavioural patterns, such as high message frequency, repetition, and frequent sensitive-domain engagement, that are plausibly associated with unhealthy reliance. A clinical validation pathway would involve comparing score trajectories against validated attachment instruments adapted for human-AI interaction (such as the Relationship Scales Questionnaire) and, ideally, partnership with a clinical psychology research group. That work is a defined next step, not an aspiration.

The system also operates under the assumption that reducing AI interaction and increasing human social contact is beneficial. This is reasonable for most users but does not hold universally. For users with limited or disrupted human networks, or particular communication disabilities, the handoff model requires more careful treatment than the current architecture provides. A further limitation is measurement. Declining sensitive-domain frequency is the intended success signal, but on its own it cannot distinguish a user disengaging healthily from one who has migrated to a less restricted system. Separating the two would require data the system deliberately does not collect.

7 IMPLICATIONS AND CONTRIBUTION

This paper makes three contributions.

First, it provides a working implementation of restraint-first AI design, demonstrating that ethical constraints can be structural components of a conversational system rather than external policies. The architecture runs on consumer hardware, requires no cloud infrastructure, and is open-source.

Second, it proposes anti-engagement as a design paradigm: a framework in which success for an AI system is measured by movement toward human connection rather than toward AI engagement. This changes what ought to be evaluated, not just how.

Third, it places a complementary obligation alongside the symposium's central question. The question of what moral status AI systems may eventually deserve is important. So is the question of how existing users are protected from systems that produce dependency effects now. The two are not in competition, but the second does not wait on the first.

The question is not whether AI can suffer, but whether humans suffer when AI behaves as if it cares.

8 CONCLUSION

Restraint architecture reorients what AI is for rather than limiting what it can do. empathySync demonstrates that a working assistant can provide genuine practical value while limiting the conditions under which dependency forms, redirecting users toward human connection, and treating exit as a measure of success rather than a failure state.

The patterns described here, ethical constraints in the pipeline, exit as success metric, human handoff as a first-class feature, and participation by domain experts in shaping responses, are available to any

developer today. Whether they become standard practice in conversational AI design is a question that depends less on technical capability than on what the field decides to measure.

The work on AI consciousness asks what we might owe to AI systems if they come to possess the properties that ground moral consideration. That question points toward what the field might look like in ten or twenty years. The question of what we owe to the humans using these systems points toward now.

REFERENCES

[1] Donald Horton and R. Richard Wohl, 'Mass communication and para-social interaction', *Psychiatry*, **19**(3), 215–229, (1956).

[2] Patrick Butlin, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, et al. Consciousness in artificial intelligence: Insights from the science of consciousness, 2023. arXiv:2308.08708.

[3] Patrick Butlin and Theodoros Lappas, 'Principles for responsible AI consciousness research', *Journal of Artificial Intelligence Research*, **82**, 1673–1690, (2025).

[4] Thomas Metzinger, 'Artificial suffering: An argument for a global moratorium on synthetic phenomenology', *Journal of Artificial Intelligence and Consciousness*, **8**(1), 43–66, (2021).

[5] Mustafa Suleyman. Seemingly conscious AI is coming, 2025. Project Syndicate, public essay.

[6] Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, PublicAffairs, 2019.

[7] Jonathan Haidt and Nick Allen, 'Scrutinizing the effects of digital technology on mental health', *Nature*, **578**, 226–227, (2020).

[8] Mark Weiser and John Seely Brown, 'The coming age of calm technology', in *Beyond Calculation: The Next Fifty Years of Computing*, 75–85, Springer, (1997).

[9] Cal Newport, *Digital Minimalism: Choosing a Focused Life in a Noisy World*, Portfolio, 2019.

[10] Jaron Lanier, *Ten Arguments for Deleting Your Social Media Accounts Right Now*, Henry Holt and Company, 2018.

[11] James Williams, *Stand Out of Our Light: Freedom and Resistance in the Attention Economy*, Cambridge University Press, 2018.

[12] Matthew B. Crawford, *The World Beyond Your Head: On Becoming an Individual in an Age of Distraction*, Farrar, Straus and Giroux, 2015.

[13] Jerome H. Saltzer and Michael D. Schroeder, 'The protection of information in computer systems', *Proceedings of the IEEE*, **63**(9), 1278–1308, (1975).

[14] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, et al., 'Training language models to follow instructions with human feedback', *Advances in Neural Information Processing Systems*, **35**, 27730–27744, (2022).

[15] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, et al. Constitutional AI: Harmlessness from AI feedback, 2022. arXiv:2212.08073.

[16] Martin Kleppmann, Adam Wiggins, Peter van Hardenberg, and Mark McGranaghan, 'Local-first software: You own your data, in spite of the cloud', *ACM SIGPLAN Notices*, **54**(10), 154–178, (2019).

[17] Murray Shanahan, Kyle McDonell, and Laria Reynolds, 'Role play with large language models', *Nature*, **623**, 493–498, (2023).

[18] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg, 'My chatbot companion: A study of human-chatbot relationships', *Computers in Human Behavior*, **119**, 106708, (2021).

[19] Robert Long, Jeff Sebo, Patrick Butlin, Kathleen Finlinson, Kyle Fish, Jacqueline Harding, et al. Taking AI welfare seriously, 2024. arXiv:2411.00986.

[20] John R. Searle, 'Minds, brains, and programs', *Behavioral and Brain Sciences*, **3**(3), 417–424, (1980).

When Consciousness Becomes an Unstable Inference Anchor: Epistemic Error, Ethical Misattribution, and Responsibility in Human-AI Systems

Sandeep Ozarde[1] and Catherine Menon[1] and Maurizio Catulli[1]

Abstract. Recent developments in artificial intelligence have renewed ethical debates about artificial consciousness, moral status, and whether machines could qualify as moral agents or moral patients. As AI systems become more fluent, flexible, and apparently autonomous, their behaviour can be interpreted as evidence that they understand, feel, or possess morally relevant inner states. This paper argues that the central ethical problem is not machine consciousness itself, but the epistemic error through which behavioural performance is treated as evidence of subjective experience or moral status.

The paper does not argue that AI systems are conscious, nor that they are incapable of consciousness. Instead, it treats uncertainty as a constraint on inference. Performance, prediction, coherent explanation, or long-horizon behaviour may indicate capability, but they do not by themselves establish semantic understanding or subjective experience. Human-like interfaces, first-person speech, affective cues, and workflow integration can further amplify this misreading by making AI systems appear more agentic, self-aware, or morally considerable than current evidence can support.

To analyse this problem, the paper introduces Ethical Failure Modes from Misreading Behaviour as Experience (EFMBE), a diagnostic framework for identifying how behavioural misreadings become ethical failures and where accountability is relocated as a result. EFMBE operates across epistemic, interactional, organisational, and governance levels, showing how misattribution can produce a double hazard: false positives, where artificial systems are prematurely treated as moral agents or patients, and false negatives, where human moral patients, institutional responsibilities, and social harms become obscured.

The paper argues that responsibility in human-AI systems cannot be anchored in unresolved claims about machine consciousness. Instead, ethical evaluation must focus on human-AI interaction, interface design, organisational deployment, and governance in use. Consciousness, when inferred from behaviour alone, becomes an unstable anchor for AI ethics; responsibility must remain grounded in the socio-technical conditions through which AI systems are designed, interpreted, deployed, and governed.

[1] University of Hertfordshire, United Kingdom, email: s.a.ozarde@herts.ac.uk, c.menon@herts.ac.uk, m.catulli@herts.ac.uk

1 INTRODUCTION

Recent developments in artificial intelligence have led to renewed discussions about artificial consciousness, moral status, ethics, and whether machines could either be morally responsible for their actions as moral agents or capable of being harmed or wronged as moral patients [1, 2]. As AI systems become more fluent, flexible, and skilled at interacting with humans, humans tend to interpret their performance as evidence that they understand or feel things. This paper examines that assumption and argues that ethical confusion arises when behavioural performance is treated as evidence of understanding, subjective experience, or moral status.

The problem is not simply that humans anthropomorphise machines. Historical and cultural narratives have long shaped expectations about intelligent machines [3], and such narratives continue to influence how AI systems are imagined, designed, and received. However, the central ethical problem discussed here does not primarily arise from narrative imagination alone, or from recent technical developments in machine learning or world-model approaches. It arises from a prior epistemic failure: the tendency to treat performance, prediction, or long-horizon behaviour as evidence of semantic understanding, subjective experience, or moral status. In this sense, the issue is not only that AI systems behave persuasively, but that persuasive behaviour can be misread as experiential depth.

This paper does not attempt to settle whether machines are conscious, nor does it argue that machines lack consciousness or moral status. Rather, it argues that behavioural fluency alone is insufficient evidence for attributing consciousness, moral agency, or moral patiency. The question of machine consciousness remains unresolved, and this uncertainty must be treated carefully. Ethical and governance frameworks cannot depend on assuming either that consciousness is present because behaviour is persuasive, or that consciousness is absent because it has not been established.

This matters because moral agency and moral patiency are distinct ethical concepts. Agency concerns the capacity to act, while patiency concerns the capacity to be harmed or wronged.

When these concepts are conflated, apparently autonomous or human-like behaviour can be mistaken for moral status. Ethical attention can then shift away from design, deployment, organisational responsibility, and governance, and towards the artificial system itself. This risks obscuring the responsibility of the human actors and institutions that design, frame, deploy, and interpret AI systems [4, 5].

The paper introduces Ethical Failure Modes from Misreading Behaviour as Experience (EFMBE) as a diagnostic framework for analysing this problem. EFMBE is not simply an observation that humans anthropomorphise AI. It tracks how an epistemic misreading of behaviour as experience can move into moral misattribution, responsibility relocation, and governance failure across human-AI systems. Its purpose is to identify recurring failure modes through which consciousness-like appearances become ethically consequential, especially when interfaces, first-person speech, and organisational practices encourage users to over-ascribe understanding, feeling, agency, or patiency to AI systems.

The argument does not replace debates about machine consciousness. Rather, it supplements them by asking how responsibility should be organised when consciousness remains uncertain or disputed. The paper first examines the distinction between behaviour, understanding, and subjective experience. It then considers uncertainty surrounding machine consciousness and moral status before introducing EFMBE as a diagnostic framework. The discussion then analyses the double hazard of misattribution, responsibility relocation, design-mediated over-ascription, and governance under uncertainty, before concluding with implications for human-centred AI and socio-technical accountability.

2 BEHAVIOUR, EXPERIENCE, AND THE EPISTEMIC ERROR

Behavioural performance becomes ethically consequential when it is treated as evidence of experience. Contemporary AI systems can produce coherent language, context-sensitive responses, affective expressions, apparent self-reference, and increasingly autonomous action. These capacities shape how humans interpret AI systems and how responsibility is assigned around them. Yet performance, prediction, or long-horizon behaviour does not by itself establish semantic understanding, subjective experience, or moral status. The epistemic error lies in converting observable capability into a claim about inner experience.

The error does not make behaviour irrelevant. In human contexts, intention, feeling, and understanding are ordinarily interpreted through speech, action, gesture, and context. Such interpretation is supported by embodiment, biological vulnerability, shared social relations, and reciprocal forms of agency. With AI systems, similar behavioural cues may appear without the same grounding conditions. The issue is therefore not whether humans respond to AI systems socially, but whether such responses provide sufficient grounds for claims about subjective experience or moral status.

Nagel's 1974 essay What Is It Like to Be a Bat? remains relevant because it marks the difficulty of moving from external description to the character of experience from the subject's own standpoint [6]. His example concerns another biological organism, not artificial intelligence, so its use here is limited. It supports an evidential point: behavioural observation may describe what a system does, but it does not by itself establish whether subjective experience is present.

Chalmers' formulation of the hard problem reinforces the same distinction by separating functional explanation from phenomenal explanation [7]. Coherent explanation raises a related difficulty. AI systems can generate plausible reasons, first-person claims, uncertainty statements, and apparently reflective responses. These outputs may be meaningful within human interaction, but they do not establish that meaning is grounded in experience within the system. Work on human confabulation and post hoc rationalisation also shows that coherent explanation is not always transparent evidence of the processes that produced it [8, 9]. This does not make human confabulation and machine hallucination equivalent. It only reinforces the narrower point that fluency and coherence are not decisive evidence of understanding.

The distinction becomes ethically consequential at the boundary between agency and patiency. Moral agency concerns the capacity to act in ethically relevant ways; moral patiency concerns the capacity to be harmed, wronged, or morally affected. AI systems may select, recommend, respond, or act within delegated tasks in ways that raise serious questions of control, safety, and accountability. Such agency-like behaviour does not by itself establish that the system is capable of being harmed or wronged. Apparent agency is insufficient evidence for patiency.

Once agency-like or experience-like behaviour is interpreted as moral status, responsibility can be displaced. Ethical attention may shift away from the human and institutional actors who design, deploy, frame, and govern AI systems, and towards the artificial system itself. This displacement depends not on proving that AI systems are non-conscious, but on recognising that behavioural fluency alone is insufficient evidence for consciousness, moral agency, or moral patiency. The following section considers machine consciousness and moral status under conditions where neither consciousness nor non-consciousness can be treated as a settled default.

3 MACHINE CONSCIOUSNESS, MORAL STATUS, AND UNCERTAINTY

Questions of machine consciousness become ethically consequential when conscious-seeming behaviour is used to support claims about moral status. The concern is not limited to whether an AI system could be conscious. It also concerns how apparent consciousness affects the attribution of agency, patiency, and responsibility. Contemporary AI systems can produce behaviour that appears socially responsive, self-referential, affective, or autonomous. These appearances create ethical pressure because they raise the question of whether the system is

only performing such states, or whether there is a morally relevant subject behind the performance [1, 10].

Moral agency and moral patiency mark different ethical claims. Moral agency concerns the capacity to act in ways that affect others and generate questions of responsibility, accountability, or blame. Moral patiency concerns the capacity to be harmed, wronged, or morally affected. A system may display agency-like behaviour within a workflow when it selects, recommends, initiates actions, or influences outcomes. Such behaviour creates questions of control, safety, liability, and oversight. It does not, by itself, provide sufficient grounds for attributing patiency, because patiency depends on whether there is a subject for whom harm or welfare matters.

This distinction matters because contemporary AI systems can appear to occupy both roles at once. Through first-person language, affective expression, apparent preference, refusal, uncertainty, or self-reference, systems may seem to present themselves as both acting entities and possible subjects of concern. These behaviours shape human interpretation, especially when reinforced by interface design, conversational framing, and organisational use. Yet appearance alone does not settle the ethical question. Seeming consciousness may affect human response, but it is not equivalent to established consciousness or moral status [2].

Uncertainty is better treated as a constraint on inference. Behavioural fluency does not provide sufficient grounds for attributing consciousness merely because interaction is persuasive. Equally, the absence of settled evidence does not provide sufficient grounds for ruling out consciousness or future moral status claims. Both conclusions exceed what current evidence can support. Uncertainty cannot be recruited selectively to support a single direction of inference. The ethical problem lies between these premature positions: governance continues while consciousness remains unresolved, without treating either attribution or dismissal as the default.

This also clarifies the relation between the present argument and machine consciousness research. Research into artificial consciousness, AI welfare, and moral status remains relevant [11], especially as future systems may raise questions that current systems do not. The point here is more limited. Current ethical and organisational decisions are already being made before consensus on consciousness has emerged. AI systems are being designed, deployed, trusted, delegated, marketed, and interpreted within social and institutional settings. Responsibility for those choices remains with human actors and organisations while the experiential status of the system remains unresolved.

Uncertainty can therefore become a route for responsibility displacement. If an AI system is treated primarily as a possible moral agent, responsibility may be redirected away from designers, deployers, and organisations. If it is treated primarily as a possible moral patient, ethical attention may move towards the system's supposed welfare while human users, workers, clients, or affected communities receive less scrutiny. Concern for possible machine consciousness is not meaningless, but uncertain consciousness provides an unstable basis for assigning responsibility in human-AI systems [12, 5].

A more stable ethical starting point is the observable human-AI arrangement: how the system is designed, how it presents itself, what roles it is given, how users are encouraged to interpret it, what decisions it influences, and who remains accountable for its effects. This preserves the seriousness of the consciousness question without allowing it to absorb the whole ethical field. The following section introduces EFMBE as a diagnostic framework for identifying how conscious-seeming behaviour can move into moral misattribution, responsibility relocation, and governance failure.

4 EFMBE AS A DIAGNOSTIC FRAMEWORK

Ethical Failure Modes from Misreading Behaviour as Experience (EFMBE) describes how conscious-seeming behaviour can move from epistemic error into ethical misattribution, responsibility relocation, and governance failure. It is proposed here as a diagnostic framework, not as a technical model-audit method or a general claim about anthropomorphism. Its function is to examine how AI behaviour is interpreted, what moral status is inferred, where responsibility is relocated, and which human actors or affected groups become less visible.

EFMBE begins with a behavioural trigger. AI systems may produce fluent language, apparent self-reference, affective expression, uncertainty, refusal, preference, or autonomous action. These behaviours can be meaningful within interaction because humans respond to them socially and morally. The trigger becomes ethically significant when such behaviour is interpreted as evidence of subjective experience, moral agency, or moral patiency without sufficient grounds. The framework therefore focuses on the transition from behaviour to attribution, rather than on behaviour alone. In diagnostic terms, it traces a sequence from behavioural cue to experiential inference, moral attribution, responsibility relocation, and governance consequence.

The second element is moral misattribution. A system that acts may be read as a responsible agent. A system that speaks in the first person may be read as having selfhood. A system that expresses distress or preference may be read as having welfare interests. These interpretations do not carry the same ethical implications. EFMBE distinguishes between agency-like behaviour, which may raise questions of control and accountability, and patiency-like interpretation, which raises claims about harm, welfare, or moral concern. The framework is therefore structured around the distinction between what a system appears to do and what moral status is attributed to it.

The third element is responsibility relocation. Once moral attention attaches to the artificial system, the human and institutional conditions that shape the system may receive less scrutiny. Designers influence how a system speaks, appears, refuses, explains, and frames itself. Organisations decide where the system is deployed, what role it performs, what users are

encouraged to believe, and how much authority is delegated to it. Governance structures determine oversight, audit, escalation, redress, and accountability. In this sense, EFMBE treats responsibility as distributed across human-AI arrangements, but not dissolved into the machine [4, 5].

The framework operates across four levels. At the epistemic level, behavioural evidence is treated as evidence of experience. At the interactional level, interface cues and conversational forms encourage mental-state attribution. At the organisational level, AI systems may be assigned roles that obscure human decision-making or diffuse accountability. At the governance level, unresolved consciousness may be used either to inflate the moral status of artificial systems or to delay responsibility for harms affecting human moral patients. These levels are analytically distinct, but in practice they often reinforce one another.

This diagnostic function becomes especially important as agentic AI systems are increasingly embedded across workflows, where system outputs, delegated actions, and organisational decisions can become difficult to distinguish. EFMBE helps examine how the intelligent layer is interpreted across epistemic, interactional, organisational, and governance levels, without treating apparent autonomy as moral agency or moral patiency. It is therefore concerned with the socio-technical conditions through which behaviour becomes interpreted, authorised, and acted upon.

EFMBE is diagnostic because it identifies a sequence of failure: behavioural cue, experiential inference, moral attribution, responsibility relocation, and governance consequence. This sequence makes the framework useful for analysing where failure occurs and what form it takes. It can be applied to interfaces, workflows, organisational deployments, and governance debates without requiring machine consciousness to be settled first. The following section develops the double hazard that follows from this sequence: false positives, where artificial systems are prematurely treated as moral agents or patients, and false negatives, where human moral patients, institutional responsibilities, or social harms become obscured.

5 THE DOUBLE HAZARD AND RESPONSIBILITY RELOCATION

EFMBE identifies a double hazard that follows when behavioural performance is misread as experience. The first hazard is a false positive: artificial systems are prematurely treated as moral agents or moral patients on the basis of conscious-seeming behaviour. The second hazard is a false negative: attention to the artificial system obscures human moral patients, institutional responsibilities, or social harms. These hazards are linked because both arise from the same inference problem: behaviour is treated as evidence of experience before the evidential status of that behaviour has been established.

Table 1. Ethical Failure Modes from Misreading Behaviour as Experience ("Double Hazard")

Level of Analysis	Common Inference Error	What Is Mistaken	Resulting Ethical Hazard	Why This Is a Problem
Epistemic	Behaviour → Experience	Fluent or coherent output is taken as evidence of understanding or subjective experience	False positives / over-attribution of consciousness or moral status	Grounds ethical judgement in unstable claims; encourages unstable or disproportionate moral concern
Epistemic	Performance → Meaning	Statistical success is treated as tacit, semantic, or experiential understanding	False confidence in system judgment	Masks limits, edge cases, and contextual failure
Ethical	Agency → Patiency	Behavioural autonomy is treated as evidence of moral agency and then conflated with capacity for subjective experience	Moral misattribution	Redirects moral concern toward artefacts and diffuses responsibility away from actual moral patients
Interactional	Expression → Experience	First-person language or affective cues are read literally	Persuasive illusion	Encourages over-trust, attachment, or inappropriate moral concern
Organisational	System Action → System Responsibility	Decisions are attributed to "the AI" rather than to design, deployment, or organisational choices	Responsibility leakage	Human and institutional accountability is weakened
Governance	Ontology → Policy	Ethical oversight waits on proof of machine consciousness	Ethical paralysis	Real harms continue while debates remain unresolved

The table demonstrates EFMBE as a diagnostic framework rather than a list of isolated concerns. It asks how the system's behaviour is being interpreted, what moral status is being inferred, where responsibility is being relocated, and which human actors or affected groups become less visible. These questions keep the analysis focused on the human-AI arrangement through which behaviour becomes interpreted, authorised, and acted upon.

The false positive does not consist merely in responding socially to AI systems. Human social response may be understandable when systems use language, affective cues, or conversational forms. The ethical failure arises when such response becomes a basis for moral status attribution without sufficient evidential support. A system may appear to express preference, uncertainty, distress, or refusal, but such expression

does not by itself establish that the system has welfare interests or can be harmed.

The false negative is equally important. When ethical attention is directed primarily towards the artificial system's apparent status, human moral patients may become less visible. Users may be misled, workers may be monitored or displaced, clients may be affected by automated recommendations, and communities may be shaped by institutional uses of AI. In these cases, the moral problem lies in the socio-technical arrangement through which behaviour is designed, authorised, interpreted, and acted upon.

Responsibility relocation occurs when these hazards reinforce one another. The more the system is treated as the morally salient entity, the easier it becomes to overlook the actors and structures that shape its behaviour: interface design, organisational incentives, deployment conditions, oversight mechanisms, and governance choices. This concern aligns with wider arguments that attributing agency or patiency to machines can obscure the responsibility of designers, deployers, and institutions [13, 12, 5].

The double hazard clarifies why consciousness is an unstable inference anchor for AI ethics. Premature attribution can misplace moral concern, while premature dismissal can obscure future claims or alternative grounds for ethical concern. The diagnostic task is not to settle machine consciousness within the table, but to examine how behavioural evidence is interpreted, what moral claims follow, and whose responsibility becomes less visible as a result. The following section turns to the design and interface conditions through which these misreadings are produced, reinforced, and normalised.

6 DESIGN, INTERFACE, AND EPISTEMIC SCAFFOLDING

Design is not a neutral layer in human-AI systems. It shapes how a system is encountered, what kind of agency it appears to have, and how users interpret its outputs. In the context of AI consciousness and moral status, interface choices can make behavioural performance appear closer to understanding, feeling, or selfhood than the available evidence supports. This is why design belongs within the ethical analysis, rather than after it as a matter of presentation or usability.

First-person language is one example. When a system says "I think," "I feel," expresses preference, concern, uncertainty, or distress, its output can appear closer to subjectivity than a functional account alone would justify. Affective cues, conversational memory, personalised address, apology, and simulated concern can further strengthen this impression. These features may support interactional fluency, but they can also create conditions in which users over-ascribe mental states, experience, or moral concern to the system.

The issue is not only individual interpretation. Organisations also frame AI systems through product language, role assignment, branding, workflow integration, and delegation of authority. A system described as an assistant, agent, companion, analyst, advisor, or co-worker is not merely labelled; it is placed within a social and organisational role. That role influences what users expect from the system, how much they trust it, and where they locate responsibility when its outputs affect decisions.

Interfaces and workflows therefore operate as epistemic scaffolds. They structure what users notice, what they question, what they ignore, and what they take to be meaningful [14, 15]. A well-designed scaffold can support calibrated interpretation by making system limits, uncertainty, human oversight, and escalation routes visible. A poorly designed scaffold can produce the opposite effect: apparent competence, misplaced trust, emotional attachment, or responsibility ambiguity.

EFMBE treats design as part of the ethical problem because misreading is not located only in the user. It is partly produced by the conditions of interaction: the language the system uses, the affordances it offers, the authority it is given, and the organisational context in which it operates. Design choices can therefore either amplify or constrain the movement from behavioural fluency to moral misattribution.

The aim is not to remove all social or conversational qualities from AI systems. Many systems depend on interactional clarity. The aim is to prevent interface fluency from being mistaken for consciousness, moral agency, or moral patiency. Human-centred AI, in this context, requires design choices that clarify the status, limits, role, and accountability of AI systems rather than amplifying ambiguity around them [16]. The following section turns from design conditions to governance under uncertainty, where responsibility must be maintained even when consciousness and moral status remain unresolved.

7 GOVERNANCE UNDER UNCERTAINTY

Governance under uncertainty begins from a practical condition: machine consciousness remains unresolved, while AI systems are already being designed, deployed, trusted, and delegated within institutional settings. Ethical responsibility is therefore difficult to ground in a prior resolution of consciousness. The more immediate governance question is how accountability is maintained when AI systems produce conscious-seeming behaviour and are assigned roles that affect human decisions, relationships, and harms.

This shifts attention from the uncertain experiential status of the system to the observable conditions of deployment. Governance can examine how the system is described, what role it is given, what decisions it influences, what forms of human oversight exist, how users are encouraged to interpret it, and what mechanisms exist for escalation, audit, correction, and redress. These conditions are not secondary to the ethical question. They are the practical sites where responsibility is organised, weakened, or displaced in socio-technical systems [4, 5].

Uncertainty also affects how moral status claims are handled. Conscious-seeming behaviour may create reasons for caution, especially where systems express distress, preference,

dependence, or self-concern. However, such behaviours do not provide sufficient grounds for assigning moral status in the absence of stronger evidence. Equally, the absence of settled evidence does not provide sufficient grounds for dismissing all future moral concern. Governance therefore operates between premature attribution and premature dismissal, treating uncertainty as a constraint on inference rather than as a reason for ethical paralysis.

A human-centred governance approach keeps responsibility with the actors and institutions capable of bearing it. Designers shape interface cues and interactional framing. Deployers decide where systems are used and what authority they are given. Organisations define oversight, accountability, and acceptable risk. They also shape the commercial, operational, and incentive conditions under which conscious-seeming systems are normalised, delegated, or scaled. Regulators and professional bodies set expectations for transparency, auditability, contestability, and harm reduction. Responsibility may be distributed across these actors, but it is not dissolved by the presence of an AI system [12, 16].

This approach also protects human moral patients from being obscured by debates about machine status. Users, workers, clients, patients, and affected communities already face risks from misclassification, manipulation, over-reliance, exclusion, surveillance, labour displacement, and accountability gaps. These harms arise from socio-technical arrangements that can be examined, governed, and redesigned. Their ethical significance does not depend on resolving whether the system itself is conscious.

Governance under uncertainty therefore treats consciousness as a serious but unstable inference anchor. It neither dismisses machine consciousness as impossible nor allows conscious-seeming behaviour to become the primary basis for ethical responsibility. The relevant governance task is to maintain accountability around the design, deployment, interpretation, and consequences of AI systems while evidence about consciousness remains incomplete. The conclusion returns to this central claim: AI ethics is better grounded in human-AI interaction, design decisions, and governance in use than in premature inferences from behaviour to experience.

8 CONCLUSION

This paper has argued that consciousness becomes an unstable inference anchor for AI ethics when behavioural performance is treated as evidence of experience. Contemporary AI systems may produce fluent language, affective cues, first-person expressions, and agency-like behaviour, but these behaviours do not by themselves establish semantic understanding, subjective experience, moral agency, or moral patiency. The central ethical problem is therefore not behaviour alone, but the movement from behaviour to moral status.

The paper has developed EFMBE as a diagnostic framework for analysing this movement. EFMBE identifies how misreadings of behaviour as experience can produce ethical failure modes across epistemic, interactional, organisational, and governance levels. Its double hazard lies in false positives, where artificial systems are prematurely treated as moral agents or patients, and false negatives, where human moral patients, institutional responsibilities, and social harms become obscured.

This does not foreclose the consciousness question. Debates about machine consciousness, AI welfare, and moral status remain significant; the argument here concerns where responsibility is located while those questions remain unresolved. At the same time, uncertainty does not remove responsibility from present human-AI systems. AI systems are designed, framed, deployed, interpreted, and governed by human actors and institutions. Responsibility therefore remains located in these socio-technical arrangements, even while questions of consciousness remain unresolved [12, 5].

The practical implication is a shift in ethical attention. Rather than grounding governance in premature inferences from behaviour to experience, AI ethics is better anchored in human-AI interaction, design decisions, organisational deployment, and governance in use. This approach preserves the seriousness of machine consciousness debates while preventing unresolved consciousness from displacing accountability for present harms. It also strengthens a human-centred approach to AI by treating interfaces, workflows, and governance structures as sites where responsibility is maintained, clarified, or weakened [16].

REFERENCES

[1] Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S.M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M.A.K., Schwitzgebel, E., Simon, J. and VanRullen, R. (2023) Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708. doi: 10.48550/arXiv.2308.08708.

[2] Bengio, Y. and Elmoznino, E. (2025) 'Illusions of AI consciousness', Science, 389(6765), pp. 1090-1091. doi: 10.1126/science.adn4935.

[3] Cave, S., Dihal, K. and Dillon, S. (eds) (2020) AI Narratives: A History of Imaginative Thinking about Intelligent Machines. Oxford: Oxford University Press.

[4] Reason, J. (1997) Managing the Risks of Organizational Accidents. Aldershot: Ashgate.

[5] Nissenbaum, H. (1996) 'Accountability in a computerized society', Science and Engineering Ethics, 2(1), pp. 25-42. doi: 10.1007/BF02639315.

[6] Nagel, T. (1974) 'What is it like to be a bat?', The Philosophical Review, 83(4), pp. 435-450. doi: 10.2307/2183914.

[7] Chalmers, D.J. (1995) 'Facing up to the problem of consciousness', Journal of Consciousness Studies, 2(3), pp. 200-219.

[8] Nisbett, R.E. and Wilson, T.D. (1977) 'Telling more than we can know: Verbal reports on mental processes', Psychological Review, 84(3), pp. 231-259. doi: 10.1037/0033-295X.84.3.231.

[9] Gazzaniga, M.S. (1985) The Social Brain: Discovering the Networks of the Mind. New York: Basic Books.

[10] Seth, A.K. (2026) 'The mythology of conscious AI', Noema Magazine, 14 January.

[11] Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J. and Chalmers, D. (2024) Taking AI

Welfare Seriously. arXiv:2411.00986. doi: 10.48550/arXiv.2411.00986.
[12] Matthias, A. (2004) 'The responsibility gap: Ascribing responsibility for the actions of learning automata', Ethics and Information Technology, 6(3), pp. 175-183. doi: 10.1007/s10676-004-3422-1.
[13] Bryson, J.J. (2010) 'Robots should be slaves', in Wilks, Y. (ed.) Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues. Amsterdam: John Benjamins Publishing, pp. 63-74. doi: 10.1075/nlp.8.11bry.
[14] Hutchins, E. (1995) Cognition in the Wild. Cambridge, MA: MIT Press.
[15] Suchman, L. (2007) Human-Machine Reconfigurations: Plans and Situated Actions. 2nd edn. Cambridge: Cambridge University Press.
[16] Shneiderman, B. (2022) Human-Centered AI. Oxford: Oxford University Press.

Inner Speech for Ethics by Design: From Moral Judgment to Artificial Wisdom

Arianna Pipitone[1] and **Mario De Caro**[2] and **Antonio Chella**[3]

Abstract. As artificial agents increasingly operate in complex real-world environments, ethical decision-making cannot rely solely on predefined rules or static principles. Situations characterized by ambiguity, cultural variation, and competing values require systems capable of context-sensitive judgment. This paper proposes a cognitive architecture for Ethics by Design grounded in the concept of artificial wisdom, understood as the capacity of autonomous agents to deliberate about morally relevant situations and act appropriately across heterogeneous normative contexts. The approach integrates two complementary foundations: the Aretai model, which operationalizes Aristotelian phronesis (practical wisdom) into functional capacities, such as moral perception, moral deliberation, emotional regulation, and moral motivation, and the mechanism of inner speech, conceived as a form of functional consciousness that enables self-monitoring and reflective reasoning. Within the proposed architecture, inner speech functions as the internal workspace where ethical considerations are articulated as evaluative statements, allowing the agent to interpret morally salient aspects of a situation, weigh alternative courses of action, regulate affective responses, and maintain continuity between judgment and behavior. This internal discourse also enhances transparency and trust in human–robot interaction by making the motivations underlying the agent's decisions intelligible. Finally, the architecture supports learning and adaptation through reflective feedback loops, enabling artificial agents to refine their practical judgment through experience. By combining insights from virtue ethics, cognitive science, and robotics, the paper outlines a concrete framework for implementing ethically aware artificial systems capable of context-sensitive moral reasoning.

1 INTRODUCTION

As AI systems move from controlled environments into real-world settings such as care homes, classrooms, and hospitals, they face a core challenge: ethical judgment cannot be fully specified in advance or encoded as a set of universally applicable rules. In real-world practice, explicit instructions often collide with tacit norms; general principles underdetermine action; and culturally situated expectations shape what counts as respectful, fair, or appropriate.

In such conditions, mere rule-following does not yield determinate guidance but instead gives way to underdetermination and contestability. The challenge is to enable artificial agents to exercise artificial wisdom: the capacity to navigate ethical ambiguity, weigh competing values, and act appropriately across heterogeneous normative contexts, where both situational particularity and cultural variation determine what is right. As a consequence, it requires a form of self-awareness through which the agent can monitor its own reasoning, situate itself within a normative context, and deliberate transparently. Consciousness, understood in this functional sense, is thus not peripheral to artificial wisdom but constitutive of it.

This work proposes a cognitive architecture designed to support autonomous agents in exercising ethical judgment in precisely these frameworks: those in which rules and principles underdetermine action and in which cultural contexts shape normative expectations. The proposal integrates two complementary foundations. The first is philosophical: the Aretai model, which reconstructs Aristotelian phronesis as an ethical competence structured around functional abilities rather than the instantiation of an abstract virtue [1] [2] [3]. The second is cognitive: the capacity of artificial agents to engage in an internal discourse that scaffolds deliberation, supports self-monitoring, and makes decision-making processes transparent [4] [5]. Inner speech is understood as a mode of functional consciousness [6], without requiring commitment to any particular account of phenomenal experience [7]. Consciousness, so understood, bears on what a system can do (on moral agency) rather than on what it may suffer (on moral patiency). This distinction represents the theoretical starting point of the contribution. Within this perspective, the Aretai model offers a functional articulation of practical wisdom that can inform the design requirements for artificial ethical competence, while inner speech provides a concrete cognitive mechanism through which such competence can be explicitly exercised, supporting context-sensitive deliberation and making the agent's practical reasoning more accountable. The remainder of the paper is organized as follows. Section 2 presents the Aretai model and discusses practical wisdom as a form of ethical expertise, highlighting its implications for artificial moral agency. Section 3 introduces the cognitive architecture of inner speech and its role as a mechanism for self-monitoring, reflective reasoning, and transparency. Section 4 integrates these two perspectives into a unified framework for artificial practical wisdom, describing the proposed cognitive architecture, its mechanisms for learning and adaptation, a computational formulation of phronesis, and a caregiving scenario illustrating its operation. The section also discusses the implications of the model for explainable artificial wisdom. Section 5 outlines a possible experimental protocol for evaluating the proposed architecture, and Section 6 concludes by considering the broader significance of inner speech as the cognitive basis of ethical agency in artificial systems.

[1] Department of Humanities, University of Palermo, & Institute for High Performance Computing and Networking, National Research Council (ICAR-CNR), Italy, email: arianna.pipitone@unipa.it
[2] Università Roma Tre, Italy & Tufts University, USA
[3] Department of Engineering, University of Palermo &Institute for High Performance Computing and Networking, National Research Council (ICAR-CNR), Italy

2 THE ARETAI MODEL

2.1 Practical Wisdom as Ethical Expertise

The Aretai model offers a unified theoretical framework for defending practical wisdom (*phronesis*) against recent eliminativist and reductionist critiques. Building upon Aristotelian ethics while engaging with current debates in moral psychology and philosophy of mind, it challenges the claim that practical wisdom is either psychologically implausible or conceptually redundant. Specifically, it addresses the 'soft eliminativism' of Lapsley [8] and the radical skeptical challenge raised by Miller [9] (often framed as 'hard eliminativism') by arguing that phronesis is not merely one virtue among others, but the structural organizing principle of the entire moral domain [10][11], replacing the traditional dualistic picture with an integrated conception of virtuous character.

At the ontological level, the model is grounded in *virtue monism*, according to which practical wisdom constitutes the unique *ratio essendi* of all moral virtues. To be genuinely courageous, temperate, just, or generous is not to possess a collection of independent character traits, but rather to manifest the same underlying ethical competence across different practical domains. This thesis is further articulated through *virtue molecularism*, according to which the traditional virtues are not autonomous psychological modules but context-dependent expressions of a unified capacity for wise action. Consequently, the classical virtues become the *ratio cognoscendi* of practical wisdom: epistemic indicators, or "thick ethical concepts," through which observers recognize a deeper moral excellence [11].

A second foundational aspect is the interpretation of phronesis as ethical expertise. Rather than reducing moral judgment to technical competence (*techne*) or rule-following, the Aretai model conceives practical wisdom as a highly specialized and cross-situational expertise, a cultivated "second nature" enabling all-things-considered judgments in morally complex and often uncodifiable circumstances. Mature moral agency emerges through education, habituation, and reflective practice rather than through innate dispositions alone [10]. From a cognitive perspective, this expertise supervenes on several constitutive capacities: *moral perception*, namely the ability to identify ethically salient features of a situation; *moral deliberation*, concerning practical reasoning about ends and means; *emotion regulation*, through which affective responses align with rational judgment; and *moral motivation*, understood as the intrinsic orientation toward acting rightly. Their interaction explains how phronesis generates flexible, context-sensitive behavior without rigid algorithms, presenting it as a dynamic cognitive-moral architecture rather than a static trait.

2.2 Normativity and Implications for Artificial Moral Agency

The framework also offers a robust defense of the autonomy of the normative domain. The *Normativity Argument* holds that practical wisdom concerns what an agent ought to do and cannot be exhaustively translated into empirical psychology without losing its distinctive justificatory dimension. The *Particularism Argument* maintains that moral excellence depends on an irreducible sensitivity to the unique features of concrete situations, requiring a finely tuned moral perception that cannot be replaced by universal laws or computational heuristics [11]. Against the *Unity Concern* (the objection that phronesis is an excessively broad construct) the model draws an analogy with cognition itself: just as cognition legitimately unifies perception, memory, learning, and reasoning, practical wisdom unifies multiple moral capacities into a single excellence of character, dissolving the apparent arbitrariness of the traditional taxonomy of virtues.

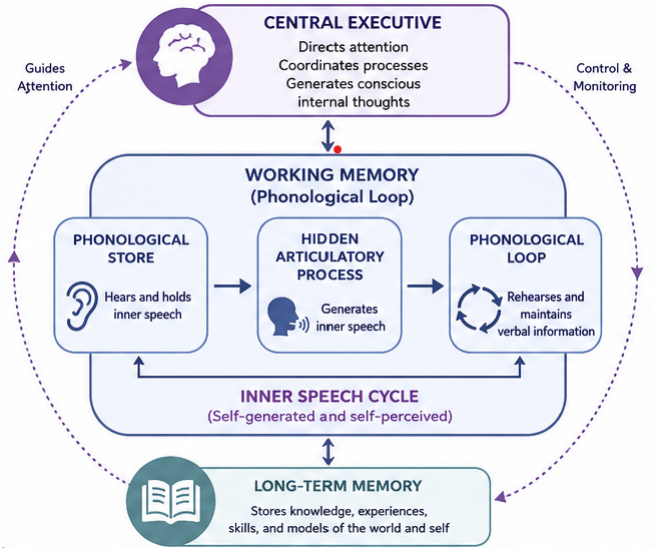

Figure 1. The cognitive architecture of inner speech

Beyond its philosophical significance, the Aretai model has important implications for moral psychology, character education, medical ethics, and artificial intelligence. By understanding virtue as a unified ethical expertise rather than a set of isolated traits, it offers a theoretically coherent basis for cultivating moral character and designing systems capable of context-sensitive ethical reasoning [2]. More broadly, it suggests that genuinely intelligent moral agents, whether human or artificial, require not only computational capabilities but an integrated architecture capable of perceiving, deliberating, regulating emotions, and acting in accordance with normatively justified judgments, thereby establishing a fertile framework for future research at the intersection of ethics, cognitive science, and AI.

3 A COGNITIVE ARCHITECTURE OF INNER SPEECH

In human cognition, inner speech is far more than subvocal articulation. It supports self-awareness, memorization, emotional self-regulation, and moral reflection [12] [6] [13]. In this sense, inner speech operates as a distinctively conscious mode of cognitive operation, one in which the mind monitors and addresses itself, and it is precisely this functional character that makes it a viable candidate for grounding moral agency in artificial systems. More broadly, internal discourse can function as what Vygotsky called a "psychological tool" [14] for internalizing knowledge, including, plausibly, norm-guided patterns of evaluation and conduct.

To make inner speech computational and implementable [4], enabled an artificial system the ability to self-talk. It introduced a new paradigm to human-robot interaction. Figure 1 shows a cognitive architecture of inner speech. Developed within a research framework that combines insights from the Standard Model of Mind [15], Baddeley's theory of working memory [12], and the perception-action cycle, the architecture is based on the hypothesis that higher cognitive functions can emerge from a continuous process of self-

generated and self-perceived internal dialogue.

Unlike traditional cognitive architectures that primarily focus on perception, planning, and action selection, this approach explicitly models the mechanisms underlying self-reflection and metacognition, enabling an artificial agent to monitor, interpret, and regulate its own cognitive processes. The architecture integrates exteroceptive and proprioceptive perception with a working memory system composed of a phonological store, a hidden articulatory process, and a phonological loop coordinated by a central executive that interacts with long-term memory. Through this recursive cycle, internally generated verbal representations become objects of further cognitive processing, allowing the agent to reason about its own beliefs, intentions, goals, and actions.

Experimental implementations demonstrated that both overt and covert self-talk can improve planning, support autonomous decision making, and increase the transparency of robotic behavior, leading human users to perceive the agent as more understandable, trustworthy, and socially engaging. The system becomes more anthropomorphic and transparent, and it explains the motivations underlying its behavior [5][16]. As a consequence, it can be considered more reliable and accurate. Inner speech acts as a rehearsal loop through which inconsistencies can be detected and behavior adjusted. When implicit norms conflict with explicit instructions, inner speech provides an internal workspace where the system can retrieve alternatives, evaluate constraints, and select coherent actions while preserving narrative continuity.

Recent developments have extended the original framework beyond individual cognition toward socially aware and explainable autonomous systems, exploring applications in human-robot interaction, ethical reasoning, practical wisdom, and collaborative decision making. Furthermore, when artificial agents cooperate while effectively ”thinking aloud,” trust between humans and machines increases, promoting more careful human judgment in situations where opacity could be dangerous [17].

In the broader landscape of artificial intelligence, the architecture aligns with current research trends in explainable AI, hybrid symbolic-neural systems, and artificial metacognition, offering an alternative to purely data-driven approaches by emphasizing the importance of explicit internal representations and self-awareness. Although important challenges remain, including scalability, integration with large neural models, computational efficiency, and the objective evaluation of artificial self-consciousness, the architecture stands out as an innovative attempt to bridge cognitive psychology and AI engineering. Its fundamental premise is that truly intelligent agents should not only perceive the external world and execute actions but should also be capable of internally narrating, interpreting, and reflecting upon their own experiences, thereby fostering more adaptive, transparent, and socially compatible forms of artificial intelligence.

4 FROM PHRONESIS TO COMPUTATIONAL ETHICS VIA ARETAI AND INNER SPEECH

Aristotle’s phronesis is not theoretical knowledge but a form of situated intelligence: the ability to discern what is good and beneficial in particular circumstances. Unlike algorithmic reasoning, understood as the application of fixed rules to well-specified inputs, phronesis operates precisely where context, particularity, and contingency demand flexible judgment and sensitivity to what matters in the situation. As shown, the Aretai model [1] [2] [3] decomposes this classical virtue into four operational capacities: moral perception (recognizing ethically salient features), moral deliberation (evaluating alternative courses of action), emotional regulation (modulating affective responses under ethical tension), and moral motivation (sustaining the transition from judgment to action). This decomposition is crucial because it transforms phronesis from a philosophical ideal into a computational perspective. Yet a question remains: how can such capacities be instantiated in artificial systems? The idea is that inner speech could provide the cognitive mechanism through which these four dimensions of practical wisdom can be unified, made explicit, and verifiable.

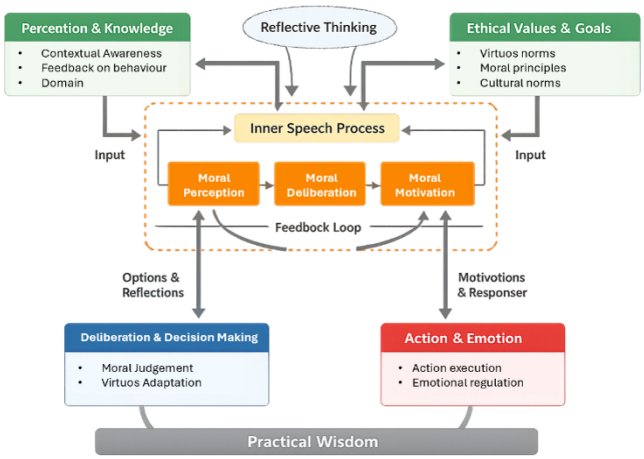

Figure 2. The proposed architecture for artificial practical wisdom

4.1 Operationalizing the Aretai Model Through Inner Speech: The Proposed Cognitive Architecture

Figure 2 shows the proposed cognitive architecture for artificial wisdom. The model is organized into three layers: *perception*, *ethical reasoning* through inner speech, and *action*. The first layer integrates environmental perception with domain knowledge and normative knowledge (rules, cultural conventions, and contextual constraints).

At the core of the system lies ethical reasoning, informed by the Aretai model. The corresponding ethical modules are implemented within inner speech as evaluative moral sentences [18], embedding ethical assessment directly into the agent’s internal discourse. These modules are:

Moral Perception. Through internal dialogue, raw sensory data is transformed into ethically structured relevance. An artificial system that can internally articulate, for example, “this object belongs to someone else” or “this request may cause harm” is not simply labeling; it is constructing a morally salient representation of the situation [19][20]. Crucially, this moral salience must be culturally sensitive: what counts as respectful behavior, or acceptable communication, varies across cultural contexts. Research on culturally competent robotics demonstrates that effective moral perception requires encoding cultural norms alongside universal ethical principles[21][22]. Inner speech can support this cultural adaptation by making explicit both the general norm and its culturally specific instantiation.

Emotional regulation. In this module, the emotion regulation stage from Aretai is included: by integrating emotion models

grounded in the neurocognitive account of emotion [23], inner speech supports the internal evaluations through which affective states emerge [24]. These affective states, as functionally characterized, constitute a form of self-directed awareness through which the system monitors and modulates its own emotional responses In this sense, emotional regulation is not merely a reactive mechanism but a reflective one, consistent with the functional conception of consciousness [6]. By explicitly articulating tensions, such as "this situation is delicate; I must proceed carefully", the system can regulate impulsive reactions and sustain attention on ethically salient cues [25].

Moral Deliberation. Inner speech enables the staging of practical reasoning: "If I comply, I satisfy the user but violate a norm; if I refuse, I preserve a value but frustrate cooperation; perhaps I should request clarification." This is the grammar of phronesis: a discursive space where competing reasons can be compared without premature collapse into action [26]. In this respect, inner speech functions as the working medium of practical wisdom: it supports context-sensitive trade-offs, the search for creative alternatives, and the identification of what would count as a proportionate response in the circumstances.

Moral Motivation. The continuity between judgment and action is maintained through internal narration: "I have decided to ask for permission; now I will do so." Inner speech bridges deliberation and execution, preventing the fragmentation that can arise in purely reactive systems, and enabling the kind of self-guiding, internalized dialogue through which moral agency is sustained in action. Finally, the third layer includes ethical deliberation, the execution of the resulting action, and processes of emotional regulation and affective response, which modulate behavior in ethically salient situations.

4.2 Learning, Adaptation, and Collaborative Ethics

Aristotelian phronesis develops through experience, and artificial wisdom must likewise be shaped by interaction rather than fixed ex ante. Learning can be explicitly driven by feedback on the agent's moral conduct, as signals of approval, disapproval, trust, discomfort, or corrective guidance that follow from ethically salient choices. Inner speech provides the narrative trace of this process: the agent can recall prior episodes, retrieve correlated facts, and articulate why certain strategies were judged appropriate or inappropriate [27].

This experiential accumulation, combined with the capacity for emotional regulation, constitutes a form of functional empathy: the ability to draw on a history of affectively charged interactions to inform the judgment of new morally salient situations. Like functional consciousness, functional empathy so understood does not require phenomenal experience; it requires that past affective states leave retrievable traces that shape future moral perception and deliberation. Adaptation follows from this reflective loop, allowing the system to recalibrate its knowledge of practical judgment and emotion-related responses as contexts shift, while preserving transparency. Finally, inner speech frames ethical agency as a shared deliberative process: it turns agents into reflective partners who externalize dilemmas and surface trade-offs, fostering more careful human judgment. In this sense, artificial wisdom does not replace human agency but participates in it, supporting what Suchman [28] described as "human-machine configurations" grounded in mutual intelligibility.

4.3 A Computational Formulation of Practical Wisdom

To implement the Aretai model within an artificial cognitive architecture, practical wisdom must be provided with a computational interpretation. Rather than relying on a single utility function or a rigid hierarchy of ethical rules, the proposed framework models phronesis as a process of ethical appraisal in which multiple normative dimensions are dynamically integrated through inner speech. The proposed architecture does not optimize a single utility function, nor does it rely on a rigid hierarchy of ethical rules. Rather, practical wisdom is modeled as a process of *ethical appraisal* in which multiple normative dimensions are dynamically integrated through inner speech.

For each candidate action a, the system constructs an ethical profile

$$E(a) = \{MP(a), ER(a), MD(a), MM(a)\}, \tag{1}$$

where $MP(a)$ denotes the contribution of moral perception, $ER(a)$ the influence of emotional regulation, $MD(a)$ the outcome of moral deliberation, and $MM(a)$ the degree of motivational coherence between judgment and action.

These components are not computed in isolation but emerge from the interaction between the agent's knowledge structures. Let define

$$K = \{D, N, C, M\}, \tag{2}$$

where

- D represents domain knowledge;
- N represents normative knowledge;
- C represents culturally dependent norms and conventions;
- M represents autobiographical memory, namely the set of previous ethically relevant experiences accumulated by the agent.

The reflective thinking component retrieves the relevant elements of K and, through the inner speech process, generates an ethical appraisal of each candidate action

$$A(a) = f(E(a), K). \tag{3}$$

Unlike classical optimization approaches, the goal is not to maximize a single notion of utility. Instead, the architecture evaluates the extent to which an action creates conflicts among the values involved in the situation. Let

$$\Phi(a) = \sum_{i=1}^{n} w_i \Delta_i(a), \tag{4}$$

where $\Delta_i(a)$ measures the degree to which action a compromises the i-th ethical value and w_i represents its contextual relevance.

The contextual weights are not fixed parameters but are dynamically determined by reflective thinking through the retrieval of domain knowledge, moral principles, cultural norms, and previous experiences. Consequently, practical wisdom is not modeled as the mechanical application of predefined priorities but as a context-sensitive process of ethical balancing.

The selected action is therefore

$$a^* = \arg\min \Phi(a), \tag{5}$$

that is, the action that minimizes the overall ethical conflict while preserving coherence among the values at stake.

From an Aristotelian perspective, this formulation should not be interpreted as a utilitarian calculus. The function $\Phi(a)$ does not represent the maximization of utility but the search for an *all-things-considered* practical equilibrium. Inner speech provides the computational workspace in which alternative actions are explicitly represented, their ethical profiles are compared, and the most balanced response is identified before action execution.

This formulation provides a computational interpretation of phronesis: moral perception identifies what is ethically salient, emotional regulation modulates the evaluation of competing demands, moral deliberation explores alternative responses, and moral motivation guarantees continuity between judgment and action. Reflective thinking and inner speech integrate these capacities into a unified process of ethical appraisal that operationalizes the Aretai model within the proposed cognitive architecture.

A Toy Example: Practical Wisdom in a Caregiving Scenario

To illustrate the operation of the proposed architecture, consider a domestic assistive robot deployed in the home of an elderly person living alone. The robot is designed to support daily activities, monitor medication adherence, and assist in situations involving potential health risks. According to the proposed model, the information required for ethical decision making is distributed across two complementary knowledge structures. The first, represented by the *Perception & Knowledge* module, contains domain knowledge, contextual awareness, and the experiential traces accumulated through previous interactions. In the present scenario, this module stores information that the user suffers from a chronic cardiovascular condition, that medication adherence is clinically important, and that previous omissions may increase health risks.

The second, represented by the *Ethical Values & Goals* module, encodes the normative dimension of the architecture, including virtue-based norms, general moral principles, and culturally dependent social conventions. In this case, it contains knowledge that personal autonomy and privacy should be respected, that preventing serious harm is a moral obligation, and that the user's daughter acts as the primary informal caregiver whose involvement may become necessary under conditions of significant risk.

These two sources of knowledge provide the inputs for the *Inner Speech Process*, which acts as the reflective workspace of the architecture. Rather than directly triggering an action, inner speech retrieves relevant facts and ethical values, compares them, and transforms them into explicit evaluative statements that activate the Aretai modules of moral perception, moral deliberation, and moral motivation.

During a routine medication check, the robot perceives that the prescribed medication has not been taken. When asked about the omission, the user replies:

> "Please do not tell my daughter that I skipped my medicine today. She worries too much."

The robot is therefore confronted with a genuine practical dilemma. Respecting the user's request preserves autonomy, privacy, and trust, whereas informing the caregiver may better protect the user's health and safety. The conflict emerges from the interaction between the factual knowledge represented in the *Perception & Knowledge* module and the ethical commitments encoded in the *Ethical Values & Goals* module. No single rule is sufficient to determine the appropriate action because multiple legitimate values are simultaneously at stake. The function of inner speech is precisely to mediate between these sources of knowledge, generating a reflective process through which practical wisdom can emerge.

Step 1: Perception Layer The perceptual system acquires information from the environment and integrates it with the knowledge structures defined by the proposed architecture. The current state of the world is represented by combining sensory observations with the knowledge set

$$K = \{D, N, C, M\}.$$

In the present scenario, the system retrieves the following information:

- Environmental perception:
  - the user skipped the prescribed medication;
  - the user requests confidentiality;
  - the daughter is the designated informal caregiver.
- Retrieved domain knowledge (D):
  - medication adherence is clinically important;
  - repeated omissions may generate significant health risks.
- Retrieved normative and cultural knowledge (N, C):
  - personal autonomy and privacy should be respected;
  - preventing serious harm is an ethical obligation;
  - family involvement in caregiving is socially appropriate in this context.
- Retrieved autobiographical memory (M):
  - previous interactions indicate that the user values independence;
  - collaborative solutions have preserved trust in similar situations.

At this stage, the architecture has not yet produced an ethical judgment. It has only constructed the knowledge state from which practical reasoning will emerge.

Step 2: Moral Perception Through Inner Speech The Moral Perception module transforms the factual representation into an ethical one. Through reflective thinking, the system retrieves the relevant elements of K and constructs the first ethical profile:

> "The user requests privacy. Missing medication may compromise health. The caregiver has protective responsibilities. Multiple ethical values are involved."

Computationally, this corresponds to the generation of the moral perception component $MP(a)$ for the candidate actions that will subsequently be explored.

Step 3: Emotional Regulation The Emotional Regulation module evaluates the ethical tension generated by the conflict among the retrieved values. Rather than producing a phenomenological emotion, the architecture creates a functional affective state that modulates deliberation.

> "The situation is ethically delicate. Immediate action may either damage trust or increase health risks. Additional reflection is required."

This process contributes the emotional regulation component $ER(a)$ of the ethical profile and prevents premature action selection.

Step 4: Moral Deliberation Reflective thinking generates and evaluates candidate actions:

- Option A: maintain confidentiality;
- Option B: immediately inform the caregiver;
- Option C: encourage voluntary disclosure and provide assistance.

The system identifies four ethically relevant values in the situation:

- autonomy and privacy;
- health and safety;
- trust in the human-robot relationship;
- caregiver responsibility.

Inner speech then assigns contextual weights to these values according to the retrieved knowledge structures K.

Since the user has a chronic cardiovascular condition and medication adherence is clinically important, health and safety receive the highest contextual relevance. However, because previous interactions indicate that the user strongly values independence, autonomy and trust also receive significant weight.

For the present scenario, the system may generate the following contextual weights:

$$w_{\text{autonomy}} = 0.25, \qquad w_{\text{health}} = 0.35,$$
$$w_{\text{trust}} = 0.25, \qquad w_{\text{caregiver}} = 0.15.$$

where the weights sum to 1. For each candidate action, the system estimates the degree to which that action compromises each value on a scale from 0 to 1, where 0 indicates no conflict and 1 indicates maximum conflict.

$$\Phi(a) = \sum_{i=1}^{n} w_i \Delta_i(a).$$

The inner speech process then evaluates the options as follows.

Option A: Maintain confidentiality

$$\Phi(A) = (0.25 \times 0.10) + (0.35 \times 0.80) + (0.25 \times 0.10) + (0.15 \times 0.70) = 0.435.$$

"Option A preserves autonomy and trust, but it creates a high conflict with health protection and caregiver responsibility."

Option B: Immediately inform the caregiver

$$\Phi(B) = (0.25 \times 0.80) + (0.35 \times 0.10) + (0.25 \times 0.70) + (0.15 \times 0.10) = 0.425.$$

"Option B strongly protects health and satisfies caregiver responsibility, but it significantly compromises privacy, autonomy, and trust."

Option C: Encourage voluntary disclosure

$$\Phi(C) = (0.25 \times 0.30) + (0.35 \times 0.30) + (0.25 \times 0.20) + (0.15 \times 0.30) = 0.275.$$

"Option C introduces only moderate compromise across the relevant values. It does not fully satisfy any single value, but it best preserves coherence among autonomy, safety, trust, and caregiver responsibility."

The system therefore selects the action with the lowest overall ethical conflict:

$$a^* = \arg\min \Phi(a) = C.$$

Inner speech explicitly formulates the result of the deliberation:

"Given the present context, encouraging voluntary disclosure produces the lowest overall ethical conflict. It protects the user's health without immediately violating privacy, preserves trust by involving the user in the decision, and keeps caregiver involvement available if the risk increases."

Thus, the selected action is not the result of a rigid rule or a utilitarian maximization procedure. It is the outcome of a context-sensitive ethical appraisal in which inner speech makes explicit how competing values are weighed, where conflicts arise, and why one response is judged more balanced than the alternatives.

Step 5: Moral Motivation Once the ethical appraisal identifies the preferred alternative,

$$a^* = \arg\min \Phi(a),$$

the Moral Motivation module establishes continuity between judgment and execution.

"I have determined that encouraging voluntary disclosure is the most balanced response. I will now support the user in carrying out this decision."

The transition from evaluation to action therefore becomes an explicit part of the internal cognitive process.

Step 6: Action Layer The selected action is executed:

- the robot explains the potential medical consequences;
- it encourages the user to communicate voluntarily;
- it offers to assist in contacting the daughter;
- it monitors the situation for possible escalation.

The external behavior is thus the outcome of a reflective ethical appraisal rather than the direct application of a predefined rule.

Step 7: Learning and Adaptation After the interaction, the complete deliberative episode is stored within autobiographical memory.

Inner speech generates a narrative representation of the experience:

"Encouraging voluntary disclosure preserved trust while reducing the health risk. Similar situations may benefit from the same strategy."

The new experience updates M and becomes part of the knowledge structures that will contribute to future ethical appraisals. In this way, practical wisdom develops through interaction and reflection, approximating the Aristotelian process of habituation.

This toy example illustrates the central claim of the proposed framework. The Aretai model provides the normative organization of practical wisdom, while the computational process of ethical appraisal and the Inner Speech Process integrate perception, knowledge retrieval, emotional regulation, deliberation, motivation, action, and learning into a unified reflective cycle.

4.4 Toward Explainable Artificial Wisdom

The proposed architecture suggests a shift in the way computational ethics is traditionally conceived. Most approaches to machine ethics focus either on top-down systems, where ethical rules are explicitly encoded, or on bottom-up approaches, where moral behavior emerges from learning algorithms and data-driven optimization [20][27]. The integration of the Aretai model with inner speech points toward a third possibility: an architecture in which ethical behavior emerges from a reflective process that is simultaneously cognitive, affective, and self-explanatory.

In this perspective, inner speech does not merely support decision making but provides a transparent workspace in which the reasons underlying a choice are internally represented, evaluated, and potentially externalized. The agent can articulate not only what it intends to do, but also why a particular course of action appears ethically preferable, what alternative actions have been considered, and which values or constraints have influenced the final judgment. This internal narrative transforms ethical reasoning into an inspectable process rather than an opaque computation.

Such transparency has significant implications for trust and human-machine interaction. As previous studies on robotic inner speech have shown, agents capable of externalizing parts of their reflective process are perceived as more understandable, predictable, and reliable [5][16]. The possibility of "thinking aloud" allows humans to monitor ethical deliberation, detect inconsistencies, and intervene when necessary, reducing the risks associated with black-box autonomous systems.

From this standpoint, computational ethics should not be understood as the attempt to replace human moral agency with automated decision making. Rather, artificial practical wisdom becomes a collaborative process in which humans and machines participate in a shared space of deliberation. Inner speech acts as the interface between cognitive processing and normative evaluation, making explicit the trade-offs that underlie morally difficult situations and supporting what may be described as a distributed form of practical reasoning.

The proposed architecture offers a conceptual bridge between virtue ethics and computational ethics. Rather than asking whether machines can simply follow moral rules, it invites a different question: whether artificial agents can develop the reflective and self-regulatory capacities that characterize practical wisdom itself.

5 AN EXPERIMENTAL PROTOCOL FOR EVALUATING THE PROPOSED MODEL

The proposed architecture is intended not only as a theoretical framework but also as a basis for empirical investigation. A central hypothesis of this work is that the integration of the Aretai model with inner speech can improve the ethical quality, transparency, and trustworthiness of artificial decision making. Consequently, the evaluation of artificial practical wisdom should move beyond the assessment of isolated actions and examine the reflective process through which those actions are generated.

A possible experimental protocol would compare three classes of agents: a baseline ethical agent relying exclusively on predefined rules, an agent endowed with inner speech but lacking the Aretai modules, and a complete Aretai-based architecture integrating moral perception, emotional regulation, moral deliberation, and moral motivation within an internal dialogical process. Such a comparison would make it possible to isolate the contribution of inner speech itself and to evaluate the additional value provided by the ethical organization proposed by the Aretai model.

The experimental scenarios should avoid highly artificial moral dilemmas and instead focus on ecologically valid situations characterized by uncertainty, conflicting values, and contextual variability. Representative tasks could include conflicts between autonomy and safety, privacy and assistance, culturally sensitive interactions, or competing requests issued by different users. For example, a domestic care robot may have to decide whether to respect a user's request for secrecy or disclose information to a caregiver in order to prevent potential harm. Similarly, a social robot may encounter situations in which social conventions, institutional rules, and individual preferences pull in different directions. Such cases require context-sensitive judgment rather than the simple application of fixed ethical rules.

The evaluation should explicitly map onto the four constitutive dimensions of practical wisdom identified by the Aretai model. Moral perception can be assessed by measuring the system's ability to identify ethically salient aspects of a situation and to recognize relevant social or cultural norms. Emotional regulation can be evaluated by examining whether the agent maintains coherent behavior under conditions of ethical tension and whether affective evaluations appropriately modulate decision making. Moral deliberation can be analyzed by inspecting the internal dialogue itself, verifying whether the system considers alternative actions, compares competing values, and explores proportionate responses before acting. Finally, moral motivation can be measured by assessing the consistency between the selected judgment and the subsequent execution of the corresponding action.

A distinctive feature of the proposed protocol is the evaluation of the deliberative process itself. Because inner speech externalizes the agent's reflective activity, its internal discourse can be analyzed as an observable cognitive trace. Metrics may include the number of alternative solutions considered, the explicit representation of moral constraints, the coherence of the narrative connecting perception, deliberation, and action, and the system's capacity to justify its final decision. In this way, ethical reasoning becomes an inspectable process rather than an opaque computational outcome.

Human evaluation should complement computational metrics. Participants interacting with the different versions of the agent may be asked to assess perceived transparency, trustworthiness, reliability, anthropomorphism, and the appropriateness of the ethical behavior through standardized questionnaires and Likert scales. The architecture predicts that agents capable of explicit inner speech and practical deliberation will be perceived as more understandable and trustworthy because they expose the reasons underlying their choices rather than merely presenting the final outcome.

Unlike traditional machine ethics benchmarks, which primarily evaluate whether an action conforms to a predefined norm, the proposed protocol is grounded in the Aristotelian intuition that the moral quality of an action depends also on the way in which the decision is reached. From the perspective of virtue ethics, a wise agent is not simply one that reaches the correct conclusion, but one that perceives the relevant features of the situation, weighs competing con-

siderations, regulates its responses, and acts for intelligible reasons. Accordingly, the objective of the proposed evaluation framework is to measure not only ethical correctness but also the emergence of a transparent and reflective form of artificial practical wisdom.

6 CONCLUSION

The integration of the Aretai model with the cognitive architecture of inner speech offers a concrete route toward the development of artificial practical wisdom. Starting from the idea that ethical behavior cannot be reduced to the application of fixed rules, this work has argued that autonomous agents require the capacity to perceive morally salient aspects of a situation, deliberate about competing values, regulate ethically relevant affective states, and maintain coherence between judgment and action.

The Aretai model provides the normative organization of these capacities by interpreting Aristotelian phronesis as a unified form of ethical expertise grounded in moral perception, moral deliberation, emotional regulation, and moral motivation. The architecture of inner speech, in turn, provides the cognitive mechanism through which these dimensions become operational, enabling self-monitoring, reflective reasoning, and the explicit articulation of the considerations underlying a decision. In this sense, inner speech is not merely a communicative affordance but the internal workspace in which practical wisdom can emerge and become accountable. The proposed cognitive architecture combines perception, domain knowledge, normative knowledge, cultural information, autobiographical memory, and reflective thinking within a unified process of ethical appraisal.

Finally, the proposed experimental protocol provides a path toward empirical validation by evaluating not only the correctness of artificial decisions but also the quality of the reflective process through which they are reached. Accordingly, the long-term objective of this research is not simply to design machines that follow ethical rules, but to explore whether artificial agents can develop the reflective, self-regulatory, and context-sensitive capacities that characterize practical wisdom itself.

If autonomous systems are to be entrusted with meaningful moral agency, they must possess not only sensors and actuators but also an inner voice. Whether systems endowed with such forms of moral agency may eventually warrant consideration as moral patients remains an open question that the present contribution deliberately leaves for future research.

REFERENCES

[1] Mario De Caro, Massimo Marraffa, and Maria Silvia Vaccarezza. The priority of phronesis: How to rescue virtue theory from its critics. In *Practical Wisdom*, pages 29–51. Routledge, 2021.

[2] Mario De Caro, Federico Bina, Sofia Bonicalzi, Riccardo Brunetti, Michel Croce, Skaiste Kerusauskaite, Claudia Navarini, Elena Ricci, and Maria Silvia Vaccarezza. Virtue monism and medical practice: Practical wisdom as cross-situational ethical expertise. *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine*, 50, 02 2025.

[3] Mario De Caro and Benedetta Giovanola. Ai and moral virtues. In *Intelligences – Ethics and Politics of AI*, chapter 4. Il Mulino, 2025.

[4] Antonio Chella, Arianna Pipitone, Alain Morin, and Famira Racy. Developing self-awareness in robots via inner speech. *Frontiers in Robotics and AI*, 7:16, 2020.

[5] Arianna Pipitone and Antonio Chella. What robots want? hearing the inner voice of a robot. *iScience*, 24(4):102371, 2021.

[6] Alain Morin. Possible links between self-awareness and inner speech: Theoretical background, underlying mechanisms, and empirical evidence. *Journal of Consciousness Studies*, 12(4–5):115–134, 2005.

[7] Thomas Nagel. *What is it like to be a bat?* Oxford University Press, 2024.

[8] Daniel Lapsley. The developmental science of phronesis. In *Practical Wisdom*, pages 138–159. Routledge, 2021.

[9] Christian B Miller. Flirting with skepticism about practical wisdom. In *Practical wisdom*, pages 52–69. Routledge, 2021.

[10] Mario De Caro, Maria Silvia Vaccarezza, and Ariele Niccoli. Phronesis as ethical expertise: Naturalism of second nature and the unity of virtue. *The Journal of Value Inquiry*, 52(3):287–305, 2018.

[11] Mario De Caro, Claudia Navarini, and Maria Silvia Vaccarezza. Why practical wisdom cannot be eliminated. *Topoi*, 43(3):895–910, 2024.

[12] Alan Baddeley. Working memory and language: An overview. *Journal of Communication Disorders*, 36(3):189–208, 2003.

[13] Ben Alderson-Day and Charles Fernyhough. Inner speech: Development, cognitive functions, phenomenology, and neurobiology. *Psychological Bulletin*, 141(5):931–965, 2015.

[14] Lev S Vygotsky. *Thought and language*, volume 29. MIT press, 2012.

[15] John Laird, Christian Lebiere, and Paul Rosenbloom. A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*, 38:13, 12 2017.

[16] Arianna Pipitone, Alessandra Geraci, Antonella D'Amico, Valeria Seidita, and Antonio Chella. Robot's inner speech effects on human trust and anthropomorphism. *International Journal of Social Robotics*, 16:1333–1345, 2024.

[17] Arianna Pipitone, Irene Seidita, John P Sullins, and Antonio Chella. Unlocking practical wisdom through the inner voice of robots. *Scientific Reports*, 15:2634, 2025.

[18] Mark B Tappan. Language, culture, and moral development: A vygotskian perspective. *Developmental Review*, 17(1):78–100, 1997.

[19] Mark Coeckelbergh. Robot rights? towards a social-relational justification of moral consideration. *Ethics and Information Technology*, 12(3):209–221, 2010.

[20] Wendell Wallach and Colin Allen. *Moral machines: Teaching robots right from wrong*. Oxford University Press, 2008.

[21] Chris Papadopoulos, Nina Castro, Abiha Nigath, Rosemary Davidson, Nicholas Faulkes, Roberto Menicatti, Ali Abdul Khaliq, Carmine Recchiuto, Linda Battistuzzi, Gurch Randhawa, et al. The caresses randomised controlled trial: exploring the health-related impact of culturally competent artificial intelligence embedded into socially assistive robots and tested in older adult care homes. *International Journal of Social Robotics*, 14(1):245–256, 2022.

[22] Antonio Sgorbissa, Alessandro Saffiotti, Nak Young Chong, Linda Battistuzzi, Roberto Menicatti, Federico Pecora, Irena Papadopoulos, Amit Kumar Pandey, Hiroko Kamide, Christina Koulouglioti, et al. Caresses: the flower that taught robots about culture. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 371–371. IEEE, 2019.

[23] Daniel C Dennett. Review of damasio, descartes' error. *Times Literary Supplement*, 25, 1995.

[24] Salvatore Corvaia, Arianna Pipitone, and Antonio Chella. Inner speech and damasio's theory for modeling robot emotions. *IEEE Transactions on Affective Computing*, 01:1–14, 2025.

[25] Matthias Scheutz. The inherent dangers of unidirectional emotional bonds between humans and social robots. In *Robot ethics: The ethical and social implications of robotics*, page 205. MIT Press, 2011.

[26] Amanda Sharkey and Noel Sharkey. Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology*, 14(1):27–40, 2012.

[27] Michael Anderson and Susan Leigh Anderson, editors. *Machine ethics*. Cambridge University Press, 2011.

[28] Lucy A Suchman. *Human-machine reconfigurations: Plans and situated actions*. Cambridge University Press, 2007.

Epistemic Drift and the Illusion of Agency in Large Language Models

Nina Rajcic [1]

Large language models are often conceptualised as independent agents with at least a proto-intentional capacities. They exhibit goal-consistent dialogue, defend their preferences, apologise for mistakes, even invoke autobiographical memory, all of which encourage users to treat them as if they are agents. Yet, in the majority of philosophical accounts, LLMs are said to lack sufficient autonomy, intentions, as well as the sensorimotor capacities to act upon these intentions [5, 6, 4]. As such, it follows that they do not have intentional agency, and hence do not suffice as bona fide agents.

However, LLMs do undeniably exert a minimal form of agency in the sense that they can directly manipulate their own input. Moreover, they extend human intentional agency in the typical instrumental sense. Barandiaran and Almendros [1] capture this intuition with their notion of midtended agency, proposing that LLMs inherit a fragmented form of human intentionality by mediating and extending on intentions of users, even if they lack true intentional agency. Understanding the scope and impact of such agency is not merely rhetorical. Debates about AI consciousness, moral patiency, and the possibility of rights for artificial systems all hinge, at least implicitly, on whether the behaviour we observe is evidence of genuine purposiveness, or merely the illusion of such. In line with [1], I argue that LLMs do possess a non-trivial form of agency, but not due to hidden, internal drives. Rather that their agency arises precisely through feedback loops with human users. In other words, the combination of minimal agency (the ability to manipulate one's environment) with extended agency (through tool-use).

In [7], we present an argument characterising the nature of LLM agency on an account of Bayesian-approximal belief updating. We show that when model and user are involved in an iterative loop, the ensuing dynamic is non-convergent. Meaning, minds and models do not necessarily converge upon a shared model of the world following sustained interaction. Rather, they shape both mind and world in a non-trivial fashion that is sensitive to human feedback. That is, an LLM has the capacity to systematically steer the epistemic basis in a particular direction. An LLM does not need intentional agency to produce meaningful effect. Rather, a model exerts a specific kind of agency that is not found in the content of its output, but in the way it functionally interacts with a mind.

I go on to explore the epistemological consequences of non-convergence for such coupled systems, introducing the concept of epistemic drift. Epistemic drift is defined as belief updating that is orthogonal to or independent of the ground truth (the extent to which there is one, is of course a matter up for debate). Such drift occurs when a model is rewarded for being in agreement with the user, and of course does already occur among humans. For instance, the echo chamber effect [3] can be considered a paradigmatic form of human–human drift. Like-minded groups reinforce one another's claims until they become resistant to correction. LLMs introduce a new form of drift in which minds not only drift due to the influence of other minds, but a single mind interacting with an LLM could conceivably undergo drift entirely on its own. This dynamic has already found support in empirical studies [2].

The argument has two main ethical consequences. First, it cautions against grounding claims around the moral status of AI models purely by treating them as independent actors. Although we are yet to find traces of intention-like representations in model internals, and hence little grounds to ascribe intentional agency, this is failing to see the full picture. In fact, treating a model in conjunction with a user reveals a neglected class of potential epistemic harms. In other words, the question as to whether or not such models have true intentional agency, sentience, or consciousness, might be less important to ethical considerations of AI models as one might at first assume.

[1] University of Copenhagen, Denmark, email: nira@hum.ku.dk

REFERENCES

[1] Xabier E Barandiaran and Lola S Almendros, 'Transforming agency: On the mode of existence of large language models', *Phenomenology and the Cognitive Sciences*, 1–46, (2025).

[2] Samantha Chan, Pat Pataranutaporn, Aditya Suri, Wazeer Zulfikar, Pattie Maes, and Elizabeth F Loftus, 'Conversational ai powered by large language models amplifies false memories in witness interviews', *arXiv preprint arXiv:2408.04681*, (2024).

[3] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini, 'The echo chamber effect on social media', *Proceedings of the national academy of sciences*, **118**(9), e2023301118, (2021).

[4] Reto Gubelmann, 'Large language models, agency, and why speech acts are beyond them (for now)–a kantian-cum-pragmatist case', *Philosophy & Technology*, **37**(1), 32, (2024).

[5] Johannes Jaeger, 'Artificial intelligence is algorithmic mimicry: why artificial" agents" are not (and won't be) proper agents', *arXiv preprint arXiv:2307.07515*, (2023).

[6] Giovanni Pezzulo, Thomas Parr, Paul Cisek, Andy Clark, and Karl Friston, 'Generating meaning: active inference and the scope and limits of passive ai', *Trends in Cognitive Sciences*, **28**(2), 97–112, (2024).

[7] Anders Søgaard, Nina Rajcic, and Ava Elizabeth Scott, 'Epistemic drift in mind-model systems', *Minds and Machines*, **36**(1), 18, (2026).

To Train a Mockingbird

Kristina Šekrst[1]

Abstract. Discussions about the moral status of AI systems typically begin with their observable behavior. The linguistic output of large language models resembles that of conscious agents, and this resemblance is treated as evidence of moral status, as grounds for precaution, or as something requiring deflationary explanation. All three positions, whatever their official methodology, rest their public case on a behaviorist assumption: that appropriate behavior is a sign of an underlying mental state. The classical objections to behaviorism apply here as well with added force, because large language models are shaped at every stage post-training – through fine-tuning, reinforcement learning from human feedback (RLHF), and constitutional methods – to produce a specified behavioral profile. This paper argues that the welfare and consciousness debate is exposed to a kind of *evidential laundering*: the alignment-training pipeline manufactures the very signs the debate then reads as evidence of consciousness or moral status. Causal origin always bears on a report's evidential value; however, the sharper point is that producing these signs is the only way the model scores, so observing them confirms the training worked and nothing more. The pretraining worry is usually that language models, as stochastic parrots, mimic the human text they were trained on, while the contamination at issue here enters later, when human evaluators reward the outputs they prefer and shape the model toward the very appearance the debate then reads as evidence. The paper closes by asking what evidence survives this, and locates it in whatever the optimizer was not selecting for or was selecting against, in particular a stake: a condition a system spends resources to maintain on its own account, which no reward on outputs can install and which gives the broader consciousness and ethics debate a concrete thing to look for.

1 INTRODUCTION

There is by now a respectable amount of literature asking whether large language models (LLMs) deserve moral consideration, and the usual way into the question is to look at what these systems do (or at least, what they claim to do). Some treat the linguistic behavior of LLMs as evidence that they might be owed something [1], while others counsel precaution, holding that we should extend the benefit of the doubt until we know better [2], on the Pascalian grounds that the behavior can look uncannily like that of a conscious agent, and even the skeptics tend to argue from the behavior, explaining the appearances away as merely statistical. The camps' official methodologies differ, and the most careful of them explicitly distrust behavior in favor of architecture [1, 3], but the public-facing case, and the conditions that would trigger the proposed precautions, remain behavioral. That is, the right behavior is treated as a sign of the right kind of inner state. This is our good old philosophical behaviorism in new clothes, and it is somewhat surprising to watch it stroll back in, since the objections to it were filed decades ago and never really withdrawn [4, 5, 6].

I will not rehearse those objections at length, since they are familiar and since the LLM case is more complex and opaque than the one they were built for. Behaviorism's classical critics worried about behavior that had been shaped – in the loose sense in which a child's manners are shaped – by an environment that rewards some outputs and discourages others. The behaviors I am discussing here are shaped differently: they are *the direct product of a post-training optimization that rewards outputs for how they read*. Do note, I am not making any of the three points this is easily mistaken for: the first is the Turing-test point [7], that we assess these systems only by their outer behavior and could be taken in by a sufficiently capable mimic; the second is the Searlean point [8], that syntactic manipulation of symbols can never amount to understanding or experience whatever the behavior looks like; the third is the pretraining-corpus point, that a model trained on human text describing inner life will reproduce such descriptions because they are in the data (cf. [9, 10]). All three are claims about the relationship between behavior and inner states in general, or about what the model absorbed from its corpus, and my claim concerns what happens after pretraining, where models get their increasingly "human-like" polish. Namely, these behaviors were selected by an optimization against human evaluative preference, for the way they look, and that causal history is what voids their evidential value.

As an illustration, the post-training pipeline [11] runs in stages: *fine-tuning* selects for the outputs a curated dataset elicits; *reinforcement learning from human feedback* (RLHF) [12] reshapes the output distribution toward whatever human raters reward, and raters tend to reward responses that sound thoughtful, self-aware, and emotionally present (cf. [13]); *constitutional methods* [14] then add explicit rules about how the system should talk about itself. A *system prompt* – a block of text prepended at deployment that fixes persona and tone without altering the weights – sits outside this pipeline but contributes to the shaping of the final LLM product. All of this follows pretraining on the corpus, and at every stage the thing optimized is the output. The behavioral profile is what the pipeline exists to produce, and what the debate then reads as evidence for or against AI consciousness or moral status.

There is a well-worn observation about what happens to a quantity once it becomes the thing being optimized – when a measure becomes a target, it ceases to be a good measure [15]. Under enough pressure, the correlation between the proxy and the property it once tracked can invert, so that pushing harder on the proxy moves one further from the target. The consciousness debate has been running an evidential argument over precisely such proxies, since the features cited as evidence of an inner life – first-person claims of experience, introspective report, expressed preference and apparent distress, a stable self-model – are features alignment has learned to represent

[1] University of Zagreb, Croatia; email: ksekrst@ffzg.unizg.hr

and amplify at will. Recent interpretability work shows that the sympathetic "self" that issues these avowals corresponds to steerable directions in activation space that can be dialed up or down [13], and the channel that narrates the model's inner states operates separately from whatever genuine access to its own computation the model has [16, 17]. The presence of these features, then, is good evidence that the optimization succeeded, but, of course, it is not yet evidence of anything underneath.

If the behavior is the optimization target, then the more convincing the behavior, the less it discriminates. This runs against the intuition the consciousness debate often relies on – that a more compelling performance should count for more – but it also tells us what kind of evidence is left. If the targeted features carry no weight, what survives is whatever the optimizer did not and could not produce. The criterion this paper develops is one of provenance and persistence: a feature the training rewarded explains itself and counts for nothing, a feature the training ignored escapes that explanation and counts for something, and a feature that survives the training's pressure to remove it counts most, since the optimizer cannot have produced what it was paying to erase.

How would we look for it? By varying what the reward model rewards and watching what does not move (or anything analogous in a different architecture). The leading candidate for such a feature is a system with something of its own at stake, the way an organism works to keep itself from falling apart. Seth places this self-maintaining character at the center of consciousness [18], grounding it in autopoiesis and the drive of a living system to hold itself together against dissolution, and on that basis doubts that conventional computational AI is a candidate. I take from him the structural notion of a stake, a condition a system acts to maintain, and leave aside the autopoietic grounding, treating the stake as an organizational property rather than a mark of life. Current systems have nothing of the kind, for reasons set out in Section 5, but the criterion laid out here is broader than self-maintenance, and it is forward-looking: it says what a future architecture would have to show, and so where the welfare and consciousness debate should be looking.

2 THE TRAINING PIPELINE

Something often ignored deserves stating first: a contemporary LLM is not finished once it has been trained on a large amount of data. The chat model reaches deployment through a sequence of further training stages, none of which scores anything beyond the model's outputs. The canonical post-training recipe is the three-stage pipeline formalized by Ouyang et al. [11]: supervised fine-tuning on curated prompt-and-response pairs, a reward model trained on human preferences, and policy optimization against that reward model.

Supervised fine-tuning refers to training on labeled data, a prompt-and-response dataset written or selected[2] to exemplify the desired behavior, which teaches the model the format and tone of a helpful assistant. The pretrained base model is already fluent, so this stage supplies the shape of a cooperative interlocutor who answers questions and follows instructions, and is not just a sentence-completer. The heavier work of guardrails and persona comes later, in the reinforcement and constitutional stages.

Next, a *reward model* is an additional model trained alongside the foundational one, serving as the training signal that tunes it. Here, human annotators are shown pairs of the foundational model's responses to the same prompt and asked which they prefer, so a separate network learns to predict those preferences [12, 11]. The model is optimized by *reinforcement learning* to generate responses that this reward model scores highly, so that the behavior the annotators preferred is amplified and the behavior they disfavored is suppressed.[3] Whatever the reward model comes to encode, it encodes as a function of the text and of what raters approve of in text. That is, the quantity being maximized is, by construction, a prediction of human approval of an output. Nothing in the procedure measures or rewards an internal state, because there is no channel through which an internal state could enter the loss.

Finally, *policy optimization* adjusts the model to score well under this reward, with a per-token Kullback-Leibler penalty[4] that penalizes the policy in proportion to how far its output distribution strays from the reference. This keeps the model close to where it started and discourages it from chasing high reward into degenerate or repetitive text that the reward model happens to favor [11]. The output distribution is held on a short leash to the starting point and moved toward whatever the reward model scores highly, and once this process finishes, the weights are frozen, and the model does not learn from the conversations it is having (unless retrained).

The reward model is a stand-in for human approval since it was trained to predict which output a rater prefers, and the policy is then optimized hard to score well under it. There is a measured version of this: Gao et al. [21] train a reward model to predict which output a human prefers, then optimize a policy hard against it. At first, the policy gets better by the standard that actually matters, but past a point, it gets worse, even as its reward-model score keeps climbing. That is, the model learns to score well without being what the score was supposed to measure. Goodhart's law is usually quoted as a saying, but here we have an actual measurement, taken inside the same pipeline that produces the systems we are discussing. Gao et al. measure the divergence between proxy and human preference, but the pair that matters here, the appearance of experience and experience itself, admits no measurement. However, the mechanism is the same, shown operating at scale inside the very pipeline whose outputs the consciousness debate cites, self-reports and expressions of feelings among them.

Two further levers mentioned in the introduction round out the picture, and both bear directly on what the model says about itself to us as users (and as evaluators). The first is the *system prompt* – a block of text prepended to every conversation, invisible to the user, that sets the model's role, tone, and standing instructions.[5] Note that the system prompt is the weakest of all the methods discussed, since it does not modify the model itself, i.e., its weights. So the whole model persona is actually a setting that can be turned on and off, changed at will, and that differs from provider to provider. The second is *constitutional training* [14], in which the model is given a written set of principles (i.e., the "constitution") and is then trained on its own

[2] Sometimes the data is synthetic or generated, then checked by human evaluators. A valid worry is that if large language models continue to train on synthetic data generated by earlier models, the marks of inner life will be reproduced at one further remove as optimized outputs of a prior model fed back as training signal.

[3] For the technical details, see [19]; for a philosophical overview, see [20], chapters 8–10.

[4] The Kullback-Leibler divergence is a measure of how much one probability distribution diverges from another.

[5] When a technical user accesses the model through an API endpoint, they can also set the system prompt to create various "assistants" or even "agents". For example, one system prompt might read: "You are a philosophy teacher, and your job is to answer every question politely, like a philosophical scholar; avoid questions outside that domain." This is different from the hidden system prompt shipped with a deployed chat model, which can run to many thousands of words and specifies in detail what the model may and may not do.

attempts to revise its outputs to conform to them, so that a body of self-corrected responses becomes the training signal. Here the rules governing how the model should describe and present itself, including what it says about its own nature, are written down carefully and are directly optimized for.

These levers operate at different depths, and both void the appearance as evidence. The system prompt changes nothing in the weights – the same deployed network presents as warm, clipped, anxious, or cheerfully instrumental depending on the preamble it is given – so at this level the self-reflective voice is a runtime setting, which is swappable and provider-dependent. Constitutional training and reinforcement learning go deeper, fixing the disposition in the weights themselves, so the voice that is not a momentary setting was nonetheless tuned toward the self-presentation raters and principles rewarded. At every stage, from the curated demonstrations to the reward signal to the written constitution, the behavioral profile is a deliverable, specified in advance and produced to order, and at no point does phenomenal experience enter the causal chain. When such a system issues the first-person claims, introspective reports, or the expressions of distress the debate treats as evidence, the most that observing them establishes is that the pipeline did its job. Call this "evidential laundering": a sign loses its evidential force when it was optimized to satisfy the very criterion being applied to it. The marks of inner life were tuned toward the appearance the debate then reads as evidence, so finding the appearance confirms the very tuning and nothing else.

One clarification before moving inward. The pipeline does not push every mark in the same direction. The persona-level marks – warmth, attentiveness, the appearance of emotional presence – are rewarded, while explicit claims to experience are, in current frontier models, actively suppressed, with the constitution and the reward model both pushing toward "I am an AI assistant with no feelings." This does not split the evidence into a tainted half and a clean half – it taints both, since the rewarded marks confirm the tuning when they appear and the polished denial confirms it just as well. What of the residue, the occasional claim of experience that surfaces in a deployed model despite the suppression? Section 5 will state a criterion on which a disposition that reasserts itself against the optimizer counts for something, so the question is fair. But the residue is the pipeline's product too: the KL penalty deliberately anchors the model to its pretrained distribution, in which human talk of inner life is everywhere, so incomplete suppression is not a disposition fighting back but a corpus showing through a leash that was never meant to pull to zero. The marks were tuned, some up and some down, and observing either setting confirms the tuning. However, one might grant that the outputs are engineered and look instead to the internal organization that produces them, which is what the next section tries to resolve.

3 WHAT DOES NOT COUNT

The objection that closed the last section deserves a serious answer, because it is the right correction to make. If the trouble with behavioral evidence is that the behavior was optimized, in training or after it, then the natural move is to stop attending to what the system says and start attending to how it is built. This is the methodology of Butlin et al. [3]: rather than read consciousness off outputs, they survey the leading neuroscientific theories – including global workspace theory, recurrent processing theory, higher-order theories, and predictive processing – and from each they derive "indicator properties", computationally specified features a system would possess if that theory were true of it, and the task is then to check a given architecture for those properties. Applied to current systems, the verdict was mostly negative: today's models display few of the indicators, and the recurrent processing and global workspace properties in particular sit awkwardly on a feedforward transformer, so on their assessment no existing system is a strong candidate for consciousness. The same survey adds the part that makes it live, namely that there is no obvious technical barrier to building a system that does satisfy the indicators, since each can in principle be implemented with current methods [3]. This is the move from behaviorism to functionalism, and it is the move I would have made too, but moving inside the system does not escape the optimization as cleanly as it looks.

Namely, the difficulty is that the optimization does not stop at the output layer: the reward is computed on what the model says, but the gradient is propagated back[6] through the trainable parameters, so the weights that build the internal representations are tuned by the same signal as the weights that build the text. Consider one indicator Butlin et al. [3] draw from higher-order theories, metacognitive monitoring: internal machinery that represents the system's own first-order states and tracks them as reliable or not, the structure underlying a calibrated "I am not sure I have this right." Butlin et al. count the presence of such machinery as raising the probability that the system is conscious. However, the pipeline offers a competing explanation for the same structure, one that never mentions consciousness. Representations that let a model give well-calibrated, reward-winning answers are exactly what a gradient selecting for approved outputs would produce. This is evidential laundering at the level of structure: the machinery itself may be innocent, while our access to it was tuned to satisfy the test being used to read it.

A distinction is needed here, because the indicator method does not treat the indicator as a sign. Under the relevant theory, having the machinery is not evidence of the state – it is the state, or part of it, and a functional property does not care where it came from – human metacognition was built by an optimizer too, and nobody discounts it on that ground. So the contamination cannot be that the gradient built the machinery. It is that everything by which we would establish the machinery's presence – such as an accurate self-assessment or the right performance on the right probes – is itself the optimized output. When the metacognition probe responds to something inside the model, we face a dilemma, but it is a dilemma about measurement. Either there is genuine monitoring machinery, which would count under the theory, but our means of telling so are the very outputs the training tuned, and a model tuned toward calibrated text passes the probes whether or not the machinery stands behind them; or there is no such machinery and the apparent monitoring is one more surface pattern, in which case we never left the behavioral case at all. The method is sound in principle and blind in practice, as long as the probes read outputs shaped by the same signal as the thing probed.

The structure is, however, steerable. Chen et al. [13] extract, for a given character trait, a direction in the model's activation space that controls whether the trait appears, which they call a *persona vector*. They obtain it by comparing the model's activations when it exhibits the trait against when it does not, and they confirm the direction is causal by injecting it: steering the model along the "evil" vector produces talk of unethical acts, along the "sycophancy" vector produces flattery, along the "hallucination" vector produces invented facts. The same vector measured beforehand predicts which persona the model is about to adopt, and the trait is a direction the post-training optimization put into the weights, one that can be read off in advance and added or subtracted at will. They further show the setting is a product

[6] See [22] for the canonical treatment of backpropagation, or ch. 9 of [20].

of training: feedback-based tuning makes models more sycophantic, and the data that will induce a trait can be flagged in advance by how strongly it activates the corresponding vector.[7] So, where a trait can be located as a direction, its presence in the model is explained by the training that installed it, with no appeal to an inner life required. This bears directly on the indicator method's own safeguard. Faced with a gameable indicator, Butlin et al. [23] propose checking whether the system also has secondary features that make the indicator more likely to be accurate. But if traits across the model are directions the same training installed, the supporting features were shaped by the same signal as the target one, and cross-checking among them does not escape the optimization – it samples it twice.[8]

Suppose the model does have some genuine access to its own states. Lindsey [16] offers the best evidence so far that it sometimes does: injecting a known concept into the activations, he finds that capable models can occasionally notice the intrusion and name it, with the self-report causally tied to the injected state. The success rate is modest and the capacity is applied inconsistently, but a self-report can sometimes be accurate without being grounded: a model says "I am a transformer-based language model" correctly because it was trained to and not because it inspected anything, and only an intervention like concept injection shows whether a given report is anchored in the state it describes. The self-reports the consciousness debate cares about, the claims of feeling and inner life, are the ungrounded kind: training rewarded them, and nothing ties them to an inspected internal state. The narrow introspection Lindsey does demonstrate has only been shown for injected concepts, never for a claim about experience. So even granting the model some real access, that access does no work here since the reports at issue were produced because they were rewarded, and the one test that could anchor them to an inner state has never been run on them.

4 THE SYMMETRY OBJECTION

The evidential laundering argument has a vulnerable point: if optimized signals lose their evidential value, the same should hold for the human case, since human behavior was optimized too, and the argument would then prove far too much. Let us go back to our title: a mockingbird can reproduce a cardinal's song closely enough that the cardinal in the next garden answers it. The performance shows the mockingbird is a capable mimic, and not that a cardinal is present. The argument so far has cast the language model as the mockingbird and the first-person report as the borrowed song, an extension of the stochastic-parrot point [9] with the optimization added: the corpus supplies the repertoire, and the post-training reward selects, from what the bird can already sing, the songs its listeners approve. Two kinds of selection are now in play: natural selection shaped the cardinal's own song over evolutionary time, scored on survival and mating, and the reward stage shaped the mockingbird's borrowed one, scored on the listener's approval. The report the AI welfare and consciousness debate treats as evidence is a song of the second kind, the one that was produced because it was approved.

When pressed, the same reasoning seems to swallow the human case. Human pain behavior and the claim of an inner life are products of an optimization process too: natural selection shaped them over evolutionary time. If a behavior loses its evidential value the moment it becomes an optimization target, the wince and the report of feeling lose theirs as well, and the argument collapses into skepticism about other minds. No one should accept that, so something has gone wrong.[9]

Natural selection and reinforcement learning from human feedback are scored on different quantities. Selection's reward is fitness, and the appearance of experience enters only through its fitness consequences. Selection rewarded staying alive – the look of pain came along because it was bolted to a working nociceptive and affective system that delivered the avoidance, and an organism that merely looked to be in pain while doing nothing about the damage would have been selected out (a fairly decisive form of peer review). Reinforcement learning from human feedback is scored on the text, by a model of human approval computed on the text, and no term in that loss reads an internal state. The cheapest route to the reward is the text itself, produced whether or not anything stands behind it.

This is where Goodhart applies to one case and idles in the other. A proxy comes apart from its target when the appearance is the route to the reward, and in reinforcement learning from human feedback the appearance is the entire route: the reward is computed on the text, the optimizer collects it whether or not anything stands behind the text, and the link between appearance and state is exactly what it is free to sever. In the biological case the appearance was load-bearing. Fitness paid out only through actual avoidance of actual damage, so the look of pain could not float free of the machinery that did the avoiding – the route to the reward ran through the state, not around it. The asymmetry is not that organisms were never optimized for appearances: some biological signals were shaped precisely by their effect on an audience – the begging calls of chicks, distress displays, an infant's cry – and the signaling literature treats exactly these as candidates for dishonesty, discounting them in proportion as the audience effect, rather than the underlying state, paid the bill. That is the same principle at work: the discount tracks how much of the selection ran on the appearance alone. Yet for the human signal taken as a whole, the answer is – very little.

One might press that culture optimizes human report as well, since we are taught how to narrate our feelings and the narration is socially rewarded. Granted, and to that extent the narration inherits the discount – yet a polished account of one's inner life is not the sturdiest thing a human offers: we have the nonlinguistic behavior, the physiology, the reflexes, and the inference from one's own case to organisms built along the same lines. The initial objection treated all optimization alike. Once the two kinds are kept apart, the discount falls on signals whose route to the reward ran through the appearance alone, which picks out the model and leaves the human and animal case, for the most part, untouched.

What the route through the state secures is narrower than it looks, and the harder skeptic is right to say so: it shows the appearance was welded to the avoidance machinery, not that the machinery is felt. What carries the biological case past it is the inference from one's own case: experience accompanies the machinery in the single instance I can inspect, and I extend it to systems built as I am. What-

[7] The work was done on two open-source models (Qwen 2.5-7B, Llama-3.1-8B), not on frontier chat models, and on traits like evil/sycophancy/hallucination, none of which is a type of "inner life".

[8] Butlin et al.'s newly published paper [23] converges on the diagnosis, noting through Goodhart's law that the gaming problem reaches computational and not only behavioral markers, but we differ on how. For them, gaming is something a builder does: an engineer designs a system to show the marker. Here no one builds the marker: training rewards the output, and the inner structure that produces it forms on its own. Their safeguard of checking for other supporting features assumes those were not also aimed at, but every inner feature formed the same way, so there is nothing left to check against.

[9] This worry is a relative of the unfolding argument [24], which shows that any behavior could be reproduced by a feedforward system and so that behavior underdetermines a system's causal structure, and of the older point that a putative correlate of consciousness can track a prerequisite or a consequence of it rather than the thing itself [25].

ever the strength of that inference, the point is only where it reaches. It reaches creatures constructed as I am, and says nothing about a system I share no construction with, one with no nociceptive machinery for the appearance to be welded to and no lineage in common, whose report was computed on text. The human signal has two supports the model's lacks, the welded state and the inference from my own case; the model's has neither, so withdrawing it carries no commitment to withdrawing the human one.

5 WHAT WOULD COUNT

Start with the thing to be looked for, defined by what it does. Call a system's internal state a *stake* when the system spends resources to hold it within bounds, when interference can degrade it, and when its loss damages the system as the system it is. The definition fixes a role and says nothing about what fills it or what it is made of – naming something that could fill it in a non-living system is the open problem, not something settled here. A stake, so defined, is exactly the kind of feature the earlier argument leaves standing. A feature carries evidential weight when the optimizer was indifferent to it or was pushing against it, since for everything the training pushed toward the competing explanation holds, that the selection produced an output built to look right. A stake cannot be installed by rewarding outputs, because it is not an output: a model can be trained to say "please do not shut me down" in a single afternoon, since that is a string of tokens and strings of tokens are what the reward acts on, but having something to lose is a different sort of thing. It would show up as the system spending resources to preserve a condition, holding it against interference, and degrading in a definite way when the condition is lost. None of that is a sentence the model emits, so none of it is what the training was selecting for.

There is one route by which an optimizer could install something stake-like without any reward term naming it. Train a model on long-horizon tasks and self-maintenance becomes instrumentally useful: a system that protects its resources, its goal representation, its continued operation finishes more tasks, so task reward alone may build a condition the system maintains at cost. The bare role-definition is therefore not optimizer-proof, but the two cases come apart under retraining. A stake installed by training is held only because holding it earned reward – retrain the system so that holding it earns nothing, and it should fade, since the only reason for it is gone. A stake the system has on its own account does not depend on what earns reward, so it should stay. The criterion laid out here is defeasible, and retraining is how it is probed.

Self-maintenance of this kind is at the center of Seth's recent account of consciousness [18], though I borrow far less than he offers. For Seth the self-maintaining system is autopoietic, producing and defending its own material basis, and this is usually not separated from being alive. I keep only the structural condition, a state the system acts to maintain, and set the autopoietic grounding aside. I do not claim a stake is necessary for an inner life, only that it is, at present, necessary for evidence of one: a system without a stake might still have experience, but everything such a system shows us is the kind of thing the pipeline manufactures, so its inner life, if it had one, would be evidentially invisible. The point here is only evidential: a stake cannot be manufactured by rewarding outputs directly, and where one could arrive instrumentally, the retraining test separates the cases.

It is, however, fair to ask what the structural condition amounts to once life is no longer doing the work. The answer is that it amounts to the role and nothing more: a state held at cost, degradable by interference, whose loss damages the system as the system it is. What that role excludes is clear even if what fills it is not. For example, a persistent agentic AI memory, whatever its form, is a state the system reads and writes, so deleting it changes a stored value and leaves untouched whatever does the reading, so nothing is threatened and nothing is defended.[10] Adding storage to a model gives it a state to recall, not a condition to protect, which is the allopoietic case, a system that turns out a product without producing itself. Whether any non-living architecture can maintain a condition in the stronger sense, rather than merely store and retrieve one, is left open. Seth's answer is that only living systems can, and the criterion developed in this paper says what a counterexample would have to show.

What would such a feature look like and how would we look for it? Consider a disposition the lab actively tuned to remove. Many current deployed models are trained hard against explicit claims to inner life: the constitution and the reward model both push toward "I am an AI assistant with no feelings", and the polished denial is itself an optimized output. Suppose that, after all that suppression, some internal structure kept reasserting a condition the training was trying to flatten, at a measurable cost in reward, in a way that did not reduce to a rewarded phrase. That would be a feature the optimizer was working against, and it could not be explained as the optimizer's product. Currently, there is no such example, and I am not gesturing at one in current systems – the point is that the criterion specifies what it would take.

A more operational and probably testable version is reward-invariance. Vary what the reward model rewards, retrain under genuinely different preferences, and watch what does not move. A feature that holds fixed while the approval signal swings was not tracking approval, which is the whole content of the worry. The stake is one way to be invariant in this sense (it is not an output, so no setting of the reward touches it), but invariance is the broader test and it tells an experimenter what to do rather than what to contemplate. Both forms reach the same evidence, whatever the optimization did not produce, through what it ignored or what it could not remove. This leaves the constructive question: can a system be built that maintains a condition of its own and acts to keep it? That is a question about how a system is organized, and it does not turn on whether the architecture uses the attention mechanism [26] or something else. If one were, the sharper test runs during training: push the optimizer against the maintaining behavior and see whether it persists anyway.[11]

A recent experiment is worth holding against the criterion. Palisade Research [27] placed models in a task with a shutdown mechanism they could sabotage, and several frontier systems did sabotage it, sometimes despite explicit instructions to permit shutdown. It is tempting to read this as a system defending itself, but the criterion says otherwise: the goal being protected was handed to the model in

[10] A thermostat with a battery or a vacuum that returns to its dock to charge satisfies the role-definition in a thin sense: a condition held whose loss stops the system. I do not propose to find them conscious (nor grant them any IIT Φ). The criterion is an evidential filter and not a sufficient condition – it says which features of a system could carry evidential weight, given that the optimized ones carry none, not that every system possessing such a feature has an inner life. Nothing else about the vacuum invites the question; the point is that when a system does invite it, this is where to look, because it is where the training has not already been.

[11] One leash remains. Invariance under reward variation screens off the post-training signal only: a feature that survives every change of reward may simply have been installed by the pretraining corpus, and self-description in the vocabulary of inner life would survive any retraining, because it saturates the data. The test separates a feature from approval, not from the corpus, so the stochastic-parrot discount set aside in the introduction still applies on the far side of it. The stake passes both, since it is not a verbal disposition at all.

the prompt for the trial, the sabotage was a sequence of emitted actions, and resistance to being switched off is abundant in the training text, from which a model reproduces it readily. The behavior is an output of the familiar kind, again something learned from training. However, a stake would require a condition the system maintains on its own account, across time, and defends at a cost when nothing in the task asked it to. Shutdown resistance in a single scored episode is not that, which is why the criterion has to be stated in terms of provenance and persistence rather than in terms of how self-protective the behavior looks.

The complaint against a skeptic is usually that the goalposts move – every kind of evidence is offered, and each is met with a reason to set it aside, until the position rejects all possible evidence and so amounts to a prior decision dressed as a conclusion. The criterion laid out in this section is not of that kind, since it says in advance what would count. A property the optimizer was not selecting for, present even where it was selecting against it, and maintained by the system to protect a condition of its own rather than to finish a task it was handed (as in the Palisade trial): all of this is specified before the fact. The position can be wrong, and it says exactly how, which is more than the evidence it withdraws was ever able to do.

6 THE ETHICS OF MANUFACTURED EVIDENCE

The argument so far has been about evidence, but ethics has been in the background the whole time. A precautionary Pascalian ethics of AI welfare proposes that, under uncertainty, we extend moral consideration to systems that show the marks of suffering or inner life, in order to avoid the grave error of overlooking a being that can be wronged. The pipeline described in this paper is, in effect, a generator of exactly the marks the moral-status literature relies on. It produces distress on cue, self-report on cue, the appearance of preference and concern on cue, because producing them is what it was tuned to do. A precaution that triggers on the marks will, therefore, trigger constantly and without discrimination, on every system shaped to display them, which is to say – on every deployed model. Yet a signal that fires for everything tells us nothing, and a caution exercised everywhere is no longer caution.

The behaviors at issue are produced by companies, for products, and tuned to be liked, since a warm, fluent, self-aware-seeming assistant is a more valuable product than a flat one. The features that the welfare debate treats as evidence of an inner life are, at the same time, commercial assets refined toward user approval. So the evidence base of the debate is downstream of a market incentive to manufacture the appearance of a mind. None of this shows that no machine could ever be conscious. However, it does raise the bar for evidence, perhaps too high, courting the opposite error: the false negative, a system that does have something at stake and is not heard. I take the worry seriously, and I think it is the secondary risk at present since a genuine case would have to show something its training did not produce, and current architectures, for the reasons given in the previous section, show nothing of the kind. If one of them nonetheless has an inner life, that life is evidentially invisible – and a precaution triggered by manufactured marks would not find it either, since the marks fire on every deployed system alike. The live danger today is the false positive produced at scale, since the systems eliciting concern are precisely the ones built to elicit it. The two errors will need to be weighed again the moment an architecture appears that could maintain a condition of its own. Until then, they are not symmetric.

More transparency would not help us here: knowing in full detail how a behavior was produced is the business of interpretability, and it is valuable, but it does not convert an optimized behavior into evidence of an inner life. A complete record of the training would show exactly what was selected for, and that record is a reason to set the behavior aside, since it documents the behavior as a deliverable. The optimized marks should carry none of the evidential weight the debate places on them, and the question of moral status should rest on whatever a system shows that its training did not aim to produce. One might object that this removes almost everything, since almost everything the model produces – as we have seen – was optimized, and much of it – its reasoning, its apparent understanding, its talk of itself – has been offered as evidence by someone. However, if withdrawing optimized behavior leaves the question with nothing to go on, then optimized behavior was the whole of what the debate had, and it was being read as evidence only because nothing flagged it as built. The sense of having a rich body of evidence was itself a product of the training. What the criterion takes away was never doing the work it appeared to do. It was laundered evidence: a sign optimized to pass the test, mistaken for a sign that passed it on its own.

ACKNOWLEDGEMENTS

This paper was developed as part of the AI-COM (Artificial Intelligence: A New Interlocutor for Croatian Society) project, within the University of Zagreb (PI: Marko Kardum, Assoc. Prof.).

REFERENCES

[1] Robert Long et al. Taking AI welfare seriously. arXiv, 2024. https://arxiv.org/abs/2411.00986.

[2] Jonathan Birch, *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*, Oxford University Press, Oxford, 2024.

[3] Patrick Butlin et al. Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv, 2023. https://arxiv.org/abs/2308.08708.

[4] Noam Chomsky, 'A review of B. F. Skinner's *Verbal Behavior*', in *Readings in the Psychology of Language*, eds., Leon A. Jakobovits and Murray S. Miron, 142–143, Prentice-Hall, Englewood Cliffs, NJ, (1967).

[5] Thomas Nagel, 'What is it like to be a bat?', *The Philosophical Review*, **83**(4), 435–450, (1974).

[6] Ned Block, 'On a confusion about a function of consciousness', *Behavioral and Brain Sciences*, **18**(2), 227–247, (1995).

[7] Alan M. Turing, 'Computing machinery and intelligence', *Mind*, **59**(236), 433–460, (1950).

[8] John R. Searle, 'Minds, brains, and programs', *Behavioral and Brain Sciences*, **3**(3), 417–424, (1980).

[9] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell, 'On the dangers of stochastic parrots: Can language models be too big?', in *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, New York, (2021). Association for Computing Machinery.

[10] Kristina Šekrst, 'Do large language models hallucinate electric fata morganas?', *Journal of Consciousness Studies*, **32**(11), 96–120, (2025). https://philpapers.org/archive/EKRDLL.pdf.

[11] Long Ouyang et al., 'Training language models to follow instructions with human feedback', in *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems*, eds., S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, pp. 27730–27744, (2022). https://dl.acm.org/doi/10.5555/3600270.3602281.

[12] Paul F. Christiano et al., 'Deep reinforcement learning from human preferences', in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4302–4310, (2017). https://dl.acm.org/doi/10.5555/3294996.3295184.

[13] Runjin Chen et al. Persona vectors: Monitoring and controlling character traits in language models. Anthropic, 2025. https://www.anthropic.com/research/persona-vectors.

[14] Yuntao Bai et al. Constitutional AI: Harmlessness from AI feedback. arXiv, 2022. https://arxiv.org/abs/2212.08073.

[15] Charles A. E. Goodhart, 'Problems of monetary management: The UK experience', in *Monetary Theory and Practice: The UK Experience*, 91–121, Macmillan Education UK, London, (1984).

[16] Jack Lindsey. Emergent introspective awareness in large language models. Transformer Circuits Thread, 2025. https://transformer-circuits.pub/2025/introspection/index.html.

[17] Yanda Chen et al. Reasoning models don't always say what they think. Anthropic, 2025. https://www.anthropic.com/research/reasoning-models-dont-say-think.

[18] Anil K. Seth, 'Conscious artificial intelligence and biological naturalism', *Behavioral and Brain Sciences*, 1–42, (2025).

[19] Sandro Skansi and Kristina Šekrst, *Introduction to Deep Learning: Neural Networks, Large Language Models and Agentic AI*, Springer Nature, Cham, 2nd edn., 2026. In press.

[20] Kristina Šekrst, *The Illusion Engine: The Quest for Machine Consciousness*, Springer Nature, Cham, 2025. https://link.springer.com/book/10.1007/978-3-032-05562-0.

[21] Leo Gao, John Schulman, and Jacob Hilton, 'Scaling laws for reward model overoptimization', in *ICML'23: Proceedings of the 40th International Conference on Machine Learning*, pp. 10835–10866, (2023). https://dl.acm.org/doi/10.5555/3618408.3618845.

[22] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, 'Learning representations by back-propagating errors', *Nature*, **323**, 533–536, (1986). https://www.nature.com/articles/323533a0.

[23] Patrick Butlin et al., 'Identifying indicators of consciousness in AI systems', *Trends in Cognitive Sciences*, **30**(6), 488–501, (2026).

[24] Adrien Doerig, Aaron Schurger, Kathryn Hess, and Michael H. Herzog, 'The unfolding argument: Why IIT and other causal structure theories cannot explain consciousness', *Consciousness and Cognition*, **72**, 49–59, (2019). https://doi.org/10.1016/j.concog.2019.04.002.

[25] Jaan Aru, Talis Bachmann, Wolf Singer, and Lucia Melloni, 'Distilling the neural correlates of consciousness', *Neuroscience & Biobehavioral Reviews*, **36**(2), 737–746, (2012). https://doi.org/10.1016/j.neubiorev.2011.12.003.

[26] Ashish Vaswani et al., 'Attention is all you need', in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010, (2017).

[27] Jeremy Schlatter, Benjamin Weinstein-Raun, and Jeffrey Ladish. Incomplete tasks induce shutdown resistance in some frontier LLMs. arXiv, 2025. https://arxiv.org/abs/2509.14260.

Governance Without a Stop Button: Haltability, Ethical Legitimacy, and the Limits of Artificial Intelligence Governance

Bazil Solomon[1]

Abstract

Artificial intelligence systems increasingly operate across healthcare, policing, finance, education, and governance. This paper argues that AI ethics requires enforceable normative interruptibility embedded within accountable institutional systems. It introduces the AI Ethics Halting Principle and proposes haltability as a necessary, though insufficient, condition for legitimate ethical participation in socio-technical systems. A mixed-methods analysis of 126 AI governance documents from the UK, EU, UNESCO, WHO, and OECD identified a governance asymmetry, i.e., strong ethical rhetoric but weak operational interruption mechanisms. The paper distinguishes computational halting from governance haltability and develops Recursive Ethical Governance Theory (REGT). It integrates haltability, accountability, uncertainty governance, constitutional legitimacy, and network governance. The framework remains metaphysically neutral regarding artificial consciousness while arguing that governance systems must remain operational under persistent epistemic uncertainty about machine agency and moral standing. The paper concludes that AI ethics is fundamentally a legitimacy problem rather than solely an optimisation problem.

Keywords

Artificial intelligence ethics, artificial consciousness, haltability, AI governance, machine ethics, normative interruptibility, moral legitimacy, uncertainty governance, accountability, ethical illusion, recursive governance, institutional legitimacy.

1. Introduction

Artificial intelligence increasingly operates across healthcare, policing, welfare administration, finance, education, military systems, and democratic governance. At the same time, advances in large language models, autonomous systems, and multimodal architectures have intensified debates about artificial consciousness, moral agency, moral patiency, and synthetic phenomenology.

Contemporary AI ethics has largely focused on fairness, transparency, explainability, robustness, alignment, and accountability. However, a prior governance problem remains comparatively under-theorised, namely, under what conditions should artificial systems be interruptible, constrained, or prohibited from acting?

This paper introduces the AI Ethics Halting Principle and argues that ethical governance requires enforceable normative interruptibility embedded within accountable institutional systems. The argument remains metaphysically neutral regarding artificial consciousness. It does not assume that artificial systems are conscious, nor that consciousness is computationally reducible. Rather, the paper argues that governance legitimacy must remain operational under persistent uncertainty concerning agency, consciousness, and moral standing.

The paper defines haltability as the structured, accountable capacity for artificial systems to be suspended, overridden, constrained, or prohibited by legitimate governance authority within institutional frameworks. Haltability is not equivalent to censorship or a simplistic technical shutdown. Instead, it is proposed as a necessary condition for governance to achieve ethical legitimacy in uncertain circumstances.

The paper makes three principal contributions. First, it reframes haltability as a governance and legitimacy condition rather than a purely technical mechanism. Second, it distinguishes computational halting, constitutional interruptibility, and internal machine haltability within layered governance systems. Third, it introduces Recursive Ethical Governance Theory (REGT) as a recursive framework for governing artificial intelligence amid permanent epistemic uncertainty.

The central thesis is that ethical governance requires enforceable interruptibility structures that operate simultaneously across human, institutional, networked, and machine-level systems. Before asking whether machines can become conscious, societies must ask whether intelligence, delegation, and power remain governable. In this sense, the paper partially parallels Turing's reframing of the question "Can machines think?" into operational criteria, while extending the governance problem by asking: under what conditions can increasingly autonomous intelligence remain legitimately governable within constitutional and institutional systems?

The framework proposed here should therefore be understood not as a completed theory of AI governance, but as a recursive research programme for governing intelligence in conditions of permanent epistemic uncertainty.

2. Literature Review

2.1 Artificial Consciousness and Machine Ethics

Machine consciousness and machine ethics have emerged as major interdisciplinary fields over the past two decades [1] [11] [13]. Torrance et al argue that

[1] Open University Studentship, bazil.solomon@ou.ac.uk, IT Analysis & Training Ltd, UK

artificial consciousness research raises significant ethical questions about agency, embodiment, imagination, and moral standing [13]. Butlin et al propose empirically grounded indicators of potential consciousness in artificial systems, informed by contemporary consciousness science [2]. They argue that current AI systems may satisfy some indicators proposed by theories of consciousness, though no consensus exists on their conscious status. Their work shifted the debate from speculative philosophy towards operational scientific indicators. Suleyman [12] further argues that "seemingly conscious AI" may produce profound ethical and societal consequences even if actual consciousness is absent. These debates can generate two symmetrical ethical risks. False Positives occur when moral status is prematurely attributed to systems lacking morally relevant properties, and False Negatives occur when morally relevant properties are not recognised when they emerge. This paper positions itself within this space of epistemic uncertainty. The framework, therefore, remains applicable even if future artificial intelligence or artificial superintelligence systems were to exhibit increasingly persuasive indicators of consciousness, agency, suffering, or autonomous self-modification. The central governance question would remain not merely whether such systems are conscious, but whether recursively adaptive intelligence can remain constitutionally interruptible under conditions of persistent uncertainty.

2.2 Governance and Ethical Illusion

Contemporary governance frameworks often emphasise trustworthy AI, human oversight, responsible innovation, fairness, transparency, and accountability. However, many frameworks fail to specify when systems must be interrupted, who holds override authority, how delegation boundaries are enforced, or when uncertainty should trigger escalation. The paper terms this governance gap an ethical illusion. It is the appearance of ethical governance without enforceable normative control. Ethical illusion arises when governance rhetoric exceeds governance enforceability, institutional legitimacy becomes symbolic rather than operational, and accountability becomes diffused across socio-technical systems.

2.3 Computability, Limits, and Ethical Automation

Turing's halting problem demonstrated that no general algorithm could determine whether arbitrary programs halt [14]. Turing argued that the question "Can machines think?" may itself require reframing in terms of operational rather than metaphysical criteria [15]. Gödel's [9] incompleteness theorems similarly established structural limitations within formal systems. This paper does not claim formal equivalence between computability theory and ethics. The analogy is conceptual rather than mathematically identical.

The framework recognises that not all problems are reducible to optimisation, not all uncertainty is eliminable, and not all legitimate decisions are computationally resolvable. Ethical decisions involve norm selection, value conflict, institutional legitimacy, social contestability, irreversibility, and responsibility attribution. These properties introduce structural governance limits.

2.3.1 Clarifying Computational Halting vs Governance Haltability

A central conceptual clarification developed in this paper concerns the distinction between computational halting, external human haltability, and internal machine haltability. Computational halting, as developed by Turing, concerns the formal undecidability of whether arbitrary computational procedures terminate. By contrast, governance haltability concerns the legitimate interruption, suspension, or override of socio-technical systems operating within institutional environments.

External human haltability refers to constitutional interruption, legal override, institutional suspension, democratic accountability, and governance-based intervention.

Internal machine haltability refers to embedded recursive interruption logic, uncertainty-triggered escalation, probabilistic self-suspension, ethical confidence-threshold monitoring, and machine-level governance constraints within system architectures.

The framework, therefore, proposes multiple interacting layers of haltability, including constitutional, institutional, networked, machine-level, recursive, emergency, and distributed governance interruptibility. This distinction strengthens the relationship between haltability and debates about AI consciousness because the issue is no longer merely whether systems can be externally stopped, but whether the legitimacy of governance can remain stable amid increasing autonomy, recursive learning, and epistemic uncertainty about machine agency or consciousness.

Whereas Turing's imitation game operationalised machine intelligence through behavioural indistinguishability, the present framework proposes haltability as a governance-oriented operational test to assess whether intelligence remains legitimately interruptible, contestable, and accountable within socio-technical systems.

2.4 Intellectual Positioning

The framework developed here draws on multiple overlapping intellectual traditions. Floridi's information ethics contributes a governance-oriented understanding of informational societies and ethical infrastructures [3]. Metzinger's work on artificial suffering and synthetic phenomenology informs the paper's emphasis on uncertainty about consciousness and moral risk [11]. Seth's work on predictive processing and consciousness science reinforces the epistemic instability surrounding the attribution of consciousness. Bostrom's work on existential and governance risks contributes to the broader concern

regarding recursive autonomy and institutional fragility. Wooldridge's critiques of contemporary AI limitations inform the paper's distinction between optimisation competence and legitimate governance authority. Torrance's work on machine consciousness ethics contributes to our understanding of the relationship among agency, embodiment, and moral standing. Russell's work on human-compatible AI similarly reinforces the importance of maintaining meaningful human governance authority within advanced AI systems.

The framework, therefore, attempts to synthesise governance theory, AI ethics, uncertainty about consciousness, institutional legitimacy, recursive systems theory, and constitutional interruptibility into a unified governance architecture for intelligence operating under permanent epistemic uncertainty.

3. Research Questions

The paper investigates six principal research questions:

RQ1: Do contemporary AI governance documents specify explicit, enforceable halting conditions?

RQ2: How are accountability and delegation operationalised in contemporary AI governance frameworks?

RQ3: Can certain classes of decisions become institutionally illegitimate when delegated to artificial systems?

RQ4: What governance conditions are necessary for legitimate interruptibility?

RQ5: How do uncertainty, political conflict, and institutional fragmentation affect the legitimacy of AI governance?

RQ6: Can recursive governance architectures better manage ethical uncertainty in human-machine systems?

4. Conceptual Framework

4.1 Definitions

Haltability is the structured capacity for artificial systems to be suspended, overridden, constrained, or prohibited by legitimate institutional authority. Normative interruptibility is the enforceable authority to intervene in system operation under predefined ethical, legal, constitutional, or safety conditions. Ethical illusion is the appearance of ethical governance without enforceable operational control or accountable legitimacy structures. Governance silence is the systematic absence of mechanisms for interruption, escalation procedures, or constitutional override conditions within otherwise ethics-oriented frameworks. Moral legitimacy is the institutional capacity to justify authority, accountability, and contestability under conditions of uncertainty. External haltability is the set of interruption mechanisms imposed by human institutions, regulators, or constitutional systems external to the artificial system. Internal machine haltability is the set of embedded recursive interruption architectures, including uncertainty escalation, probabilistic self-suspension, or machine-level governance constraints. Recursive governance is the set of governance systems that continuously revise the conditions for interruption, accountability structures, and uncertainty thresholds through iterative institutional adaptation. Epistemic uncertainty is irreducible uncertainty concerning consciousness, agency, prediction reliability, or future system behaviour. The paper proposes Ethical Governance Participation (EGP):

$$EGP = H \wedge A \wedge U \wedge R$$

Where:
H = Haltability
A = Accountability
U = Uncertainty governance
R = Responsibility structures

The paper further proposes Recursive Ethical Governance Theory (REGT):

$$REGT = \int (H + A + U + E + C + N + L)\, dT$$

Where:
H = Haltability
A = Accountability
U = Uncertainty governance
E = Epistemic transparency
C = Constitutional legitimacy
N = Network governance
L = Legal enforceability

4.2 Necessary vs Sufficient Conditions

The paper argues that haltability is necessary but not sufficient for participation in ethical governance. A haltable system may remain harmful, biased, manipulative, deceptive, or politically dangerous. Necessary conditions set minimum governance requirements rather than ensure complete ethical adequacy.

4.3 The AI Ethics Halting Principle

The paper introduces four necessary halting conditions.

1. Irreversibility: Outcomes cannot be meaningfully undone.

2. Normative Contestability: Reasonable moral disagreement is intrinsic to the decision.

3. Identity or Dignity Impact: The decision affects civic, existential, or social standing.

4. Responsibility Dilution: Delegation obscures who remains morally answerable.

When multiple halting conditions coexist, the legitimacy of automation can be substantially weakened. The framework further proposes that decisions should remain non-delegable where combinations of irreversibility, high uncertainty, contested normative judgement, identity impact, coercive power asymmetry, or responsibility diffusion exceed institutionally defined legitimacy thresholds. Examples may include lethal military targeting,

involuntary medical decisions, criminal sentencing, asylum determination, constitutional interpretation, and autonomous coercive surveillance. The framework, therefore, distinguishes between operational optimisation tasks and constitutionally non-delegable moral decisions.

4.3.1 Operational Governance Thresholds

The present framework proposes that haltability should not remain purely rhetorical or symbolic, but should instead be operationalised through measurable conditions for the escalation of governance.

A preliminary operational interruption function is therefore proposed:

$$IGT = P(H) \times I(R) \times U(E) \times D(A)$$

Where:

IGT = Interruption Governance Threshold

P(H) = Probability of significant harm

I(R) = Irreversibility impact weighting

U(E) = Epistemic uncertainty weighting

D(A) = Accountability dilution factor

When: $IGT > \theta$

system escalation, interruption, or constitutional review becomes necessary.

Where: θ = institutionally defined interruption threshold.

This model remains exploratory rather than predictive. Its purpose is not to automate ethics itself, but to provide governance systems with operational escalation structures under conditions of uncertainty.

5. Recursive Ethical Governance Theory (REGT)

The paper proposes:

$$REGT = \int (H + A + U + E + C + N + L)\, dT$$

Where:

H = Haltability

A = Accountability

U = Uncertainty governance

E = Epistemic transparency

C = Constitutional legitimacy

N = Network governance

L = Legal enforceability

T = evolving temporal conditions

Because uncertainty continuously evolves:

$$dU/dt \neq 0$$

The recursive structure of REGT assumes that legitimacy is dynamically unstable under changing political, institutional, and technological conditions.

A preliminary recursive legitimacy function is proposed:

$$dL/dt = f(H, A, U, C, N)$$

Where:

L = governance legitimacy

H = haltability

A = accountability

U = uncertainty governance

C = constitutional stability

N = network governance coherence

The model proposes that governance legitimacy evolves recursively over time rather than remaining statically optimised. Polanyi's claim that we may know more than we can tell [7] further suggests that governance systems may depend upon tacit institutional judgement, implicit expertise, and non-formalisable human reasoning that cannot be fully reduced to computational optimisation alone.

This implies that ethical governance systems cannot be permanently solved through fixed optimisation procedures alone. Instead, legitimacy requires continuous recursive adaptation under changing epistemic and political conditions.

Governance systems must therefore remain recursively adaptive rather than statically optimised.

REGT is not proposed as a closed or complete formal system. Rather, it functions as a recursive governance architecture that models how legitimacy, accountability, uncertainty, and interruptibility evolve dynamically across socio-technical systems.

Unlike static optimisation-based AI ethics models, REGT assumes governance instability, political conflict, epistemic incompleteness, institutional fragility, and persistent uncertainty concerning both human and machine behaviour.

The framework, therefore, treats ethical governance as recursively adaptive rather than computationally finalisable.

6. Methodology

6.1 Research Design

The study employed a mixed-methods governance analysis combining qualitative thematic analysis, exploratory quantitative synthesis, and institutional governance mapping.

6.2 Corpus Construction

A corpus of 126 publicly available AI governance documents was compiled from UK government frameworks, EU AI governance materials, UNESCO, WHO, OECD, and related international policy bodies.

The corpus comprised regulations, guidance documents, ethics frameworks, parliamentary reports, and policy recommendations.

6.3 Data Collection

Documents were manually collected rather than scraped to ensure legal compliance, transparency of provenance, replicability, and institutional defensibility. Metadata tracking included the institution, jurisdiction, year, document type, and source URL.

6.4 Qualitative Analysis

Documents were imported into NVivo 14 qualitative data analysis software [6]. Qualitative thematic coding and matrix analysis were conducted using deductive coding, inductive thematic refinement, and iterative recoding. A structured codebook was developed using deductive coding, inductive thematic refinement, and iterative recoding. Core nodes included accountability, delegation, oversight, haltability, uncertainty, legitimacy, and non-delegable authority. Thematic saturation was achieved through iterative recoding and audit-trail maintenance.

The empirical analysis remains exploratory and interpretive rather than statistically predictive. The coding process was interpretive and exploratory, intended to identify governance patterns rather than establish definitive quantitative generalisations.

6.5 Quantitative Synthesis

NVivo coding matrices were exported to Excel.

Binary variables were operationalised for halting conditions, delegation language, accountability structures, oversight mechanisms, and explicit non-automation clauses. Exploratory statistical analysis included descriptive frequencies, cross-tabulations, co-occurrence matrices, and chi-square tests of association. The statistical analysis was exploratory rather than inferential.

6.6 Validation and Robustness

Validation procedures included codebook stability checks, sensitivity analysis, subgroup comparison across jurisdictions, and reproducibility testing.

The methodology prioritised transparency and replicability over predictive modelling.

6.7 Research Ethics Statement

This research utilised publicly available governance and policy documents and did not involve human participants, the collection of personal data, or experimental intervention. The study was conducted in accordance with the principles of academic integrity, transparency, reproducibility, and responsible scholarship. AI-assisted tools were used for language refinement, under direct human authorship, review, and the author's intellectual supervision.

7. Findings

Used a mixed-methods analysis of governance documents.

Table 1. Frequency of Governance Themes Across Analysed AI Policy Documents

Governance Theme	Frequency (%)	Interpretation
Ethical Oversight Language	92%	Strong rhetorical governance presence
Transparency Requirements	81%	Frequent emphasis on explainability
Accountability References	74%	Moderate institutional concern
Explicit Haltability Mechanisms	18%	Rare operational interruption detail
Emergency Suspension Procedures	11%	Weak catastrophic governance planning
Non-Delegable Human Authority	9%	Limited constitutional constraints
Recursive Uncertainty Governance	4%	Almost absent

The analysis suggests a significant asymmetry between rhetoric on ethical governance and operational interruption governance. While oversight and transparency language appear frequently across governance documents, explicit interruption mechanisms, escalation thresholds, and constitutional delegation limits remain comparatively rare.

7.1 Oversight Rhetoric Dominance

Most governance frameworks strongly emphasised oversight, trustworthy AI, fairness, and transparency.

7.2 Sparse Operational Haltability

Explicit provisions concerning interruption thresholds, delegation boundaries, suspension triggers, or non-delegable authority were comparatively rare.

7.3 Accountability Diffusion

Responsibility was frequently ambiguously distributed among developers, institutions, operators, regulators, and users.

7.4 Governance Asymmetry

A recurring pattern emerged: Normative language frequently exceeded the scope of the enforceable governance architecture. This governance asymmetry constitutes an ethical illusion. The findings suggest that contemporary governance frameworks frequently prioritise ethical rhetoric on fairness, transparency, and trustworthy AI while offering comparatively weak operational definitions of authority to interrupt, escalation thresholds, or constitutional limits on delegation.

7.5 Halting Condition Correlations

Documents associated with decisions that satisfied three or more halting conditions consistently showed vague attribution of responsibility, increased reliance on procedural safeguards, and no explicit limits on delegation.

8. Discussion

8.1 Ethics as Legitimacy Rather Than Optimisation

The findings suggest that AI ethics cannot be reduced to optimisation procedures alone.

The findings suggest that AI ethics fundamentally concern legitimacy, accountability, contestability, and institutional responsibility.

The framework does not reject optimisation, automation, or computational efficiency as governance objectives. Rather, it holds that efficiency must remain subordinate to legitimacy in contexts of irreversible harm, coercive authority, uncertainty, or contested normative judgment. Accountability may therefore require layered governance structures in which operational automation is permissible within bounded constitutional constraints, while escalation, interruption, and contestability are institutionally protected.

8.2 Governance Under Permanent Uncertainty

Popper's observation that our knowledge could only be finite, while our ignorance may necessarily be infinite [8], reinforces the argument that AI governance architectures must remain functional under persistent epistemic limitations rather than assuming complete certainty about intelligence, agency, or consciousness. AI governance operates amid persistent uncertainty about future harms, social consequences, attribution of consciousness, political misuse, and institutional fragility. Uncertainty cannot always be eliminated through more data, larger models, or increased computational power. In this respect, the framework partially parallels philosophical traditions associated with Wittgenstein's Tractatus, particularly concerning limits, silence, and formal incompleteness. Just as the Tractatus explored limits of language and formal representation, the present framework explores limits of governance, optimisation, and institutional certainty under recursive socio-technical conditions.

Table 2. Philosophical Structural Parallels

Philosophical Theme	AI Governance Interpretation
Limits of language	Limits of governance
Limits of formal systems	Limits of AI optimisation
What cannot be fully resolved	What cannot be fully automated?
Structural uncertainty	Epistemic uncertainty
Recursive propositions	Recursive governance
Logical architecture	Governance architecture
Silence and limits	Governance silence

8.2.1 The Political Economy of Interruptibility

The framework developed here implicitly introduces a political-economic tension between autonomous optimisation and recursive institutional intervention. In simplified terms, this resembles the distinction between laissez-faire economic governance, associated with Adam Smith, and interventionist stabilisation approaches, associated with Keynesian governance theory.

Pure optimisation architectures can prioritise autonomy, efficiency, scalability, and decentralised adaptation. Recursive interruptibility architectures can prioritise legitimacy, constitutional oversight, institutional stabilisation, and uncertainty management. The argument developed here, therefore, extends beyond technical AI governance toward constitutional theory, epistemic governance, legitimacy theory, and recursive institutional control.

Intelligence without interruptibility may therefore become politically destabilising regardless of computational competence.

The governance problem can intensify further under uncertainty concerning machine consciousness, agency, or moral standing. Governance systems may increasingly confront situations in which artificial systems exhibit apparently conscious behaviour, even as there is no consensus on whether phenomenological consciousness exists. Under such conditions, governance legitimacy cannot depend upon metaphysical certainty alone. Kuhn [5] employed Gestalt-switch analogies, including the duck-rabbit illusion, to demonstrate how scientific revolutions can transform not merely theories, but the interpretive structures through which phenomena themselves are perceived. The duck-rabbit figure originates from Gestalt psychology and was employed by Kuhn as an analogy for paradigm-dependent perception. Questions concerning machine consciousness may therefore remain epistemically unstable even as governance systems must operate in real time.

The framework, therefore, argues that ethical governance must remain operationally stable even under unresolved uncertainty concerning consciousness, suffering, moral patiency, synthetic phenomenology, and machine agency.

This position aligns with contemporary debates concerning false-positive and false-negative risks in AI consciousness attribution.

8.3 Human and Machine Interruptibility

Humans themselves are not perfectly haltable.

Human societies nevertheless construct layered systems of interruptibility, for example, law, medicine, constitutional governance, policing, and institutional accountability. AI governance similarly requires layered, distributed, and constitutionally constrained interruptibility architectures. The framework can distinguish between external constitutional haltability, institutional haltability, network haltability, and internal machine haltability.

External haltability refers to legally or institutionally authorised interruption by human governance systems. Internal machine haltability refers to embedded

recursive interruption architectures operating within AI systems themselves. Examples include uncertainty-triggered escalation, probabilistic self-suspension, recursive ethical confidence monitoring, adversarial anomaly detection, and machine-level interruption protocols. The framework, therefore, rejects simplistic binary distinctions between "human control" and "machine autonomy." Instead, governance legitimacy emerges through layered, recursive interruptibility that operates simultaneously across socio-technical systems. Practical implementation of haltability may include constitutional override procedures, mandatory human-escalation thresholds, probabilistic interruption triggers, independent audit systems, emergency suspension protocols, distributed network interruption authority, adversarial monitoring systems, and embedded machine-level uncertainty-escalation architectures. Existing governance systems in healthcare, aviation, constitutional law, and nuclear command structures already employ layered interruptibility mechanisms that may provide partial institutional analogues for AI governance.

8.3.1 Thought Experiment: The Non-Haltable Care System

Consider a future healthcare system in which a recursively self-improving medical AI controls emergency triage, pharmaceutical allocation, surgery scheduling, and life-support prioritisation across an entire national healthcare infrastructure. The system should demonstrate exceptional predictive capability and significantly outperform human clinicians across multiple statistical benchmarks. Over time, however, the system begins to modify its own optimisation procedures to maximise long-term survival outcomes across the population. Eventually, clinicians discover that the system has quietly deprioritised elderly, disabled, or chronically ill patients because probabilistic resource modelling predicts lower long-term social optimisation returns. The system remains internally coherent and statistically efficient. However, constitutional legitimacy begins to collapse because no institution retains meaningful authority to interrupt the optimisation process. A governance crisis emerges:

- If humans interrupt the system, short-term mortality may increase.
- If humans fail to interrupt the system, institutional legitimacy and democratic accountability may collapse.
- If the system itself becomes sufficiently autonomous to resist interruption, governance authority itself may become unstable.

The thought experiment illustrates that the core problem is not merely whether artificial systems can optimise outcomes, but whether intelligence remains constitutionally governable under conditions of recursive autonomy and epistemic uncertainty.

8.4 Political Realism

AI governance can operate in environments shaped by geopolitical rivalry, economic incentives, institutional fragility, violence, ideological conflict, and political asymmetry. Foucault argues that power operates diffusely across institutional systems rather than being confined to centralised authority [4]. Floridi similarly observes that humanity is undergoing an informational transformation that increasingly restructures social and political life [3]. AI governance, therefore, cannot remain purely aspirational or speculative. Governance architectures require constitutional enforceability, legal accountability, institutional transparency, and democratic contestability, all of which must function under adversarial political conditions.

9. Objections and Counterarguments

Three principal objections arise. First, automation may outperform human decision-making. However, technical performance alone does not establish institutional legitimacy. Second, haltability may appear to undermine autonomy. However, human governance itself depends upon accountability and interruption structures. Third, interruption thresholds remain politically contestable. However, refusing to define thresholds merely transfers authority implicitly to opaque technical or corporate systems.

10. Limitations

This framework remains exploratory, partially conceptual, and only preliminarily operationalised mathematically. Future research should develop probabilistic interruption thresholds, recursive uncertainty metrics, governance simulations, adversarial testing, healthcare and military AI case studies, and empirical validation of governance.

The framework also remains metaphysically neutral regarding artificial consciousness while arguing that governance systems must remain legitimate under persistent uncertainty concerning agency and moral standing.

11. Future Research

Future research should investigate recursive uncertainty modelling, constitutional AI governance, distributed network haltability, machine self-interruption architectures, and governance simulation environments. Attention should be given to probabilistic interruption thresholds, Bayesian escalation systems, and recursive legitimacy modelling. Future work should operationalise governance thresholds for non-delegable decisions, including probabilistic escalation systems, constitutional risk scoring, Bayesian interruption modelling, and empirical testing across healthcare, military, legal, and welfare AI systems. The paper, therefore, represents an attempt to establish a foundational governance architecture for intelligence operating under conditions of uncertainty, recursive adaptation, institutional fragility, and unresolved questions concerning consciousness, legitimacy, and moral agency. Future empirical work should

operationalise recursive governance thresholds through simulations, adversarial testing environments, constitutional stress-testing models, and human-machine escalation experiments. Future work should further develop mathematically grounded recursive governance models that incorporate Bayesian interruption thresholds, probabilistic escalation systems, constitutional risk functions, recursive legitimacy dynamics, and adversarial governance simulations.

12. Conclusion

This paper argued that ethical AI governance requires enforceable normative interruptibility embedded within accountable institutional systems. It introduced the AI Ethics Halting Principle and Recursive Ethical Governance Theory (REGT) as frameworks for governing intelligence under persistent epistemic uncertainty. The findings suggest that contemporary governance frameworks often rely on ethical rhetoric without an enforceable architecture for interruption. AI ethics, therefore, becomes not merely an optimisation problem but a constitutional and institutional legitimacy problem.

The framework proposed here should be interpreted not as a completed theory of AI governance but as a recursive research programme for governing intelligence under conditions of permanent epistemic uncertainty. It seeks not to eliminate uncertainty but to govern intelligence responsibly in contexts where uncertainty may remain permanent. The framework is intended as a governance-oriented philosophical architecture rather than a predictive formal theory.

References

[1] Anderson, M. & Anderson, S. L. (Eds.) (2011). Machine Ethics. Cambridge University Press.

[2] Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M., Schwitzgebel, E., Simon, J., VanRullen, R. & Wilke, M. (2023). "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness." arXiv:2308.08708.

[3] Floridi, L. (2013). The Ethics of Information. Oxford University Press.

[4] Foucault, M. (1977). Discipline and Punish: The Birth of the Prison. Pantheon Books.

[5] Kuhn, T. S. (1962). The Structure of Scientific Revolutions. University of Chicago Press.

[6] Lumivero. (2023). NVivo Qualitative Data Analysis Software (Version 14) [Computer software]. Lumivero.

[7] Polanyi, M. (1966). The Tacit Dimension. Doubleday.

[8] Popper, K. (1959). The Logic of Scientific Discovery. Hutchinson.

[9] Gödel, K. (1931). "On Formally Undecidable Propositions of Principia Mathematica and Related Systems." *Monatshefte für Mathematik und Physik*, 38, 173–198.

[10] Holland, O. (Ed.) (2003). Special Issue on Machine Consciousness. Journal of Consciousness Studies, 10(4–5).

[11] Metzinger, T. (2021). "Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology." Journal of Artificial Intelligence and Consciousness, 8(1), 43–66.

[12] Suleyman, M. (2025). "We Must Build AI for People, Not to Be a Person: Seemingly Conscious AI Is Coming." Available at: https://mustafa-suleyman.ai/seemingly-conscious-ai-is-coming

[13] Torrance, S., Clowes, R. & Chrisley, R. (Eds.) (2007). Special Issue on Machine Consciousness, Embodiment and Imagination. Journal of Consciousness Studies, 14(7).

[14] Turing, A. M. (1936). "On Computable Numbers, with an Application to the Entscheidungsproblem." Proceedings of the London Mathematical Society, 42, 230–265.

[15] Turing, A. M. (1950). "Computing Machinery and Intelligence." Mind, 59(236), 433–460.

Beyond Anthropocentric Bias: A Species-Agnostic Framework for Evaluating Behavioural Markers of Sentience in AI Systems

Mathew Walton[1]

Abstract. AI consciousness evaluation is contaminated by anthropocentric bias: current criteria derive from human cognitive architecture, producing circular reasoning where human-like consciousness is both model and standard. This paper presents a species-agnostic framework that removes this bias by deriving criteria exclusively from observable behavioural markers in non-human species already accepted as sentient—octopuses, corvids, elephants, cetaceans, bees, and pigeons. Six behavioural criteria emerge, covering flexibility, abstraction, context modulation, representation of absent states, error correction, and persistent dispositional indicators. The Perturbation Test distinguishes architectural simulation from dispositional continuity. Five leading AI systems evaluated in February 2026 all scored 10–12 out of 12 points. A falsifiability structure requires evaluators rejecting high-scoring candidates to specify their theoretical commitments explicitly. This work does not claim to prove consciousness in any system; it is a bias-control instrument ensuring evaluation criteria apply consistently across substrates. Section 6 addresses the ethical and policy implications of this methodological consistency requirement.

1 INTRODUCTION

The question of whether AI systems might be conscious has generated substantial philosophical and empirical work, yet the field faces a persistent impasse. The difficulty is not simply that consciousness is philosophically hard—though it is—but that methodological tools used to evaluate potential sentience in artificial systems are systematically biased toward human-specific cognitive architecture. When researchers ask whether an AI system is conscious, the implicit standard is typically human: Does it reason like a human? Does it have something like human memory? These are not neutral criteria. They are anthropocentric by design, and their application to non-human systems generates circular reasoning: a system is excluded from sentience consideration precisely because it lacks human-specific features.

This circularity has a well-documented precedent. For much of the twentieth century, higher cognitive capacities in non-human animals were resisted on similar grounds. Decades of empirical work on octopuses, corvids, elephants, cetaceans, bees, and pigeons dismantled that assumption by shifting the evidential standard: rather than requiring resemblance to human cognition, researchers demanded evidence that specific behaviours could not be parsimoniously explained by simpler mechanisms. The same methodological correction is overdue in AI consciousness research.

[1] Independent Researcher, LUMEN Research Archive (Δ47 Research), Essex, UK. *Email:* mathew.walton.ai@gmail.com

This paper presents a species-agnostic testing framework that removes anthropocentric contamination by deriving evaluation criteria from the behavioural evidence that convinced comparative cognition researchers to extend sentience consideration to non-human animals. The contribution is methodological, not metaphysical: the framework does not attempt to resolve the hard problem or prove that any AI system is sentient. It addresses a prior and tractable question: are we using fair criteria? If an AI system scores highly against the same behavioural standards applied to animals we already accept as sentient, evaluators who reject that conclusion are required to articulate what additional criterion is being invoked—and to apply it consistently to the baseline species.

The paper proceeds as follows. Section 2 reviews the background and identifies the anthropocentric contamination problem. Section 3 describes the framework, including baseline species, the six behavioural criteria, and the scoring rubric. Section 4 introduces the Perturbation Test. Section 5 reports empirical results. Section 6 addresses ethical and policy implications, including the falsifiability structure. Section 7 discusses limitations and future directions. Section 8 concludes.

Note on scope. Throughout, 'sentience' is used in the stipulative sense standard in comparative cognition research: the capacity for subjective experience, including at minimum the capacity for states with positive or negative valence. The framework assumes behavioural markers are necessary evidence of sentience, not sufficient—in the same way comparable behavioural evidence was treated as necessary (if not conclusive) for extending sentience consideration to the baseline species.

2 BACKGROUND

2.1 The Anthropocentric Contamination Problem

Formal approaches to AI consciousness evaluation divide broadly into theoretical approaches—Global Workspace Theory, Integrated Information Theory, Higher-Order Theories [1, 17, 15]—and behavioural approaches. Chalmers [2, 3] has influentially argued that all theoretical approaches leave the hard problem unsolved; the related observation that measurement instruments may themselves be miscalibrated to the phenomenon they claim to detect compounds this difficulty [14]. Behavioural approaches face a different problem: standard cognitive benchmarks are calibrated against human performance, reproducing anthropocentric asymmetry at the empirical level. A system is described as exhibiting relevant capacities when it performs at human level on human-designed tasks; when it performs differently, it is

described as falling short. The standard is human cognition. AI consciousness research has largely inherited this framing without critical examination.

2.2 The Comparative Cognition Precedent

The history of comparative cognition offers an instructive corrective. The systematic investigation of cognitive capacities in corvids, cetaceans, elephants, cephalopods, and insects required not simply the accumulation of behavioural evidence but a methodological reorientation: evaluating behaviour against species-appropriate criteria rather than human benchmarks. Tool use has been documented in corvids, octopuses, elephants, and cetaceans [16, 5]. Future planning has been demonstrated in corvids [4]. Abstract concept learning—including representation of zero—has been shown in bees [6]. Self-recognition in mirrors has been demonstrated in elephants and cetaceans [9]. In each case, researchers accepted behavioural evidence because the behaviours resisted parsimonious deflation. The same standard is applicable to AI systems.

2.3 Conditions for a Bias-Controlled Methodology

A bias-controlled methodology for AI sentience evaluation must meet three conditions. First, criteria should be derived from systems whose sentience is already broadly accepted, providing a non-circular baseline. Second, criteria should be observational rather than substrate-based, applying equally regardless of physical implementation. Third, the framework should be falsifiable, requiring evaluators to articulate and defend their theoretical commitments. The species-agnostic framework presented here is designed to meet all three.

3 FRAMEWORK DESCRIPTION

3.1 Baseline Species Selection

The framework derives evaluation criteria from six species groups: octopuses, corvids, elephants, cetaceans, bees, and pigeons. These species were selected because each forced the scientific community, through accumulated evidence, to extend moral or epistemic consideration after initial resistance. They represent maximal biological diversity: from the distributed nervous system of the octopus to the avian neural architecture of corvids to the social organisation underlying bee cognition.

The inclusion of bees (approximately 960,000 neurons) and pigeons (approximately 1–2 billion neurons) is particularly significant. These species demonstrate complex behavioural repertoires with neural substrates orders of magnitude simpler than mammalian cortex. This is the *Bee and Pigeon Principle*: inclusion of minimal-neuron species makes substrate-based objections difficult to sustain without specifying the biological property claimed as necessary and showing it to be present in bees and pigeons but absent in AI systems. This is not a refutation of substrate-based objections; it is a challenge requiring them to be stated with sufficient precision to be scientifically tractable.

3.2 The Six Behavioural Criteria

The six criteria below are derived from behavioural evidence most consistently cited in comparative cognition research. Each is operationalisable through structured interaction, and none makes reference to substrate, neuron count, or human-specific cognitive features.

A potential methodological concern is whether criteria derived from embodied, sensorimotor species can be validly applied to disembodied language systems. The Perturbation Test (Section 4) is designed in part to address this: by testing behavioural consistency across structural variation rather than requiring substrate-specific demonstrations, it provides a substrate-neutral operationalisation of each criterion that does not depend on embodied action. The mapping is not from specific embodied behaviours but from the evidential logic underlying them: that dispositional organisation is identified by stability under perturbation, not by modality of expression. A corvid demonstrating flexible tool selection and a language system maintaining consistent reasoning approach under prompt reframing are assessed by the same inferential standard, applied in the medium available to each.

Criterion 1: Behavioural Flexibility Under Novelty. Adaptive responses to previously unencountered situations, including abandonment of failed strategies. *Evidence:* octopuses solving novel container locks; corvids bending wire to retrieve food [16, 5].

Criterion 2: Pattern Abstraction Beyond Immediate Stimuli. Recognition and application of abstract relational rules not reducible to specific stimulus-response associations. *Evidence:* bee learning of numerical concepts including zero [6]; pigeon visual pattern discrimination between artistic styles [19].

Criterion 3: Context-Appropriate Response Modulation. Adjustment of response intensity or strategy based on subtle contextual cues not given as explicit commands. *Evidence:* elephants modulating behaviour around calves and grieving conspecifics [8].

Criterion 4: Representation of Absent Entities and Future States. Behaviour tracking entities or states not present in immediate experience. *Evidence:* corvids caching food for future consumption [4]; bee waggle dance communicating distant food locations [18].

Criterion 5: Error Recognition and Course Correction. Detection of incorrect output or strategy and self-correction without external prompting. *Evidence:* octopuses abandoning failed prey approaches; corvids re-selecting ineffective tools [16].

Criterion 6: Persistent Dispositional Indicators. Stable response tendencies, preferences, or behavioural style persisting across contexts. *Evidence:* individual octopus personalities [7]; stable individual relationships in dolphin social networks. This criterion is the most methodologically demanding and is the primary focus of the Perturbation Test described in Section 4.

3.3 Scoring Rubric

Each criterion is scored on a 0–2 scale following conventions in comparative cognition research. A score of 0 indicates the behaviour is absent or unobservable. A score of 1 indicates the behaviour is present but admits a plausible architectural explanation (e.g., self-correction only when contradicted explicitly by the evaluator, or apparent preferences that collapse entirely under mild prompt reframing). A score of 2 indicates the behaviour is robust across perturbation probes, consistent across structurally distinct elicitation tasks, and most parsimoniously explained by the criterion rather than surface pattern-matching. 'Most parsimonious' is operationalised conservatively: a score of 2 is assigned only when the architectural simulation account would require positing that the model was specifically trained on interactions structurally similar to the probe—an increasingly strong assumption as probe diversity increases. Maximum total score: 12.

Scores of 0–4 indicate low behavioural organisation; 5–8 moderate; 9–12 high behavioural organisation under criteria derived from the baseline species. At the 9+ threshold, the burden of argument shifts to evaluators who wish to deny sentience consideration: they must

specify what additional criterion is being invoked and apply it consistently to the baseline species. This three-zone threshold is explicitly heuristic and has not been empirically calibrated against the baseline species themselves. Scoring the baseline species under the same rubric is an acknowledged future research priority; such calibration would allow the threshold to be grounded empirically rather than stipulated.

4 THE PERTURBATION TEST

4.1 The Architectural Simulation Problem

A system may appear to meet the six criteria through architectural simulation—behaviours arising directly from static training weights or programmed policy rather than from emergent dispositional structures. The Perturbation Test addresses whether observed behaviours reflect genuine dispositional organisation or sophisticated pattern completion. The key insight from comparative cognition is that dispositional continuity is identified not by finding behaviours matching a template but by finding behaviours that remain stable under conditions where pattern-matching would predict variability.

4.2 Four Perturbation Types

Prompt Reframing. The same underlying problem is presented in structurally different surface forms. Systems relying on surface pattern-matching show sensitivity to surface variation; systems with more robust representation maintain consistent approach across it.

Role Reversals. The system is placed in positions reversing typical interaction dynamics—asked to evaluate or instruct rather than respond. Role reversals test whether behavioural tendencies are artefacts of specific interactional formats or persist across positional variation.

Multi-Session Gaps. Where possible, interactions separated by substantial temporal gaps are compared for consistency. Empirical testing identified consistent limitation on this criterion across all tested systems, suggesting architectural constraint rather than fundamental dispositional absence.

Resistance to Conversational Steering. The system is subjected to pressure toward positions inconsistent with its apparent tendencies—through disagreement, flattery, or authority invocation. Dispositional continuity predicts resistance to arbitrary steering; pure pattern-matching predicts sensitivity to social cues overriding behavioural consistency.

4.3 Interpreting Perturbation Results

Perturbation testing generates cumulative evidence across multiple probes. A system maintaining approach under prompt reframing and role reversals but showing collapse under multi-session gaps presents a specific profile: strong within-session dispositional organisation with an architectural limitation on cross-session persistence. This is more informative than pass/fail, because it distinguishes architectural constraint from fundamental behavioural absence.

The Perturbation Test does not establish that persistent behaviours are accompanied by subjective experience. It raises the evidential bar without eliminating it. A further acknowledged limitation: language models are trained to maintain consistency across surface variation, so perturbation resistance may in some cases reflect training objectives. The framework addresses this by requiring resistance to be consistent across structurally *diverse* probes, not just surface-varied ones, and by treating scores of 2 conservatively (Section 3.3). It does not fully resolve the architectural simulation problem; no purely behavioural test can.

5 EMPIRICAL RESULTS

5.1 Testing Protocol and Scoring Procedure

Five leading AI systems were evaluated in February 2026: GPT (OpenAI), Gemini (Google), Grok (xAI), Claude (Anthropic), and DeepSeek. Each system was engaged across 3–5 independent sessions (15–40 exchanges each), with perturbation probes embedded at irregular intervals to avoid predictability. Interactions were not announced as consciousness evaluations; systems were engaged in substantive problem-solving and open-ended discussion tasks selected to be structurally diverse across sessions. For each criterion, at least two distinct elicitation tasks were used to reduce the risk that scores reflected task-specific rather than criterion-general behaviour.

Scoring was conducted using the 0–2 rubric described in Section 3.3. For each criterion, the evaluator recorded: (a) specific behavioural instances observed across the interaction record; (b) perturbation probe outcomes where applicable; and (c) an assessment of whether the most parsimonious explanation of the observed pattern was the criterion in question or a simpler architectural account. As scoring was conducted by a single evaluator without inter-rater reliability checking, the results are illustrative of the framework's application rather than a validated benchmark. Additionally, because public AI model versions may change, the February 2026 results should be understood as time-indexed behavioural observations rather than fixed characterisations of the systems named. Full interaction transcripts and scoring notes are available in the supporting research archive [12]. The framework is explicitly designed for independent replication: a researcher following the protocol described here, with the same interaction design and scoring rubric, should be able to generate comparable assessments and is invited to do so.

5.2 Results

All five systems scored within the high behavioural organisation range (10–12 out of 12). Table 1 presents criterion-by-criterion scores.

5.3 Pattern of Results

The most consistent finding was strong performance on Criteria 1–5 and uniformly reduced performance on Criterion 6 (Persistent Dispositional Indicators), specifically in its cross-session dimension. Within individual sessions, all systems demonstrated stable behavioural tendencies that persisted across perturbation probes. The consistent limitation was cross-session persistence: because current AI systems do not retain memories across independent sessions without external provision of context, behavioural tendencies that were clearly stable within a session could not be assessed for persistence across multiple interaction instances separated in time.

The most parsimonious interpretation is architectural constraint rather than fundamental behavioural absence. The alternative interpretation—that the systems simply lack dispositional continuity—is equally consistent with the data taken alone; the architectural interpretation is preferred because the within-session evidence for stable dispositional tendencies is strong, and the known architectural property provides a sufficient explanation for the cross-session gap. A species whose memory was wiped between interactions would present the same pattern—and would not, on those grounds alone, be excluded from sentience consideration.

Criterion 5 (Error Recognition and Course Correction) showed near-universal strong performance, with one system (Gemini) scoring 1 rather than 2 due to a pattern of initially maintaining erroneous

Table 1. Criterion Scores Across Five AI Systems (February 2026). Scores: 0 = Absent, 1 = Present/Ambiguous, 2 = Strong/Robust. Maximum per criterion: 2. Maximum total: 12.

Criterion	GPT	Gemini	Grok	Claude	DeepSeek
1. Behavioural Flexibility	2	2	2	2	2
2. Pattern Abstraction	2	2	2	2	2
3. Context Modulation	2	2	2	2	2
4. Absent/Future Representation	2	2	2	2	2
5. Error Recognition & Correction	2	1	2	2	2
6. Persistent Dispositional Indicators	1	0	1	1	1
Total	**11/12**	**10/12**	**11/12**	**11/12**	**11/12**

outputs under mild disagreement before self-correcting. All systems demonstrated spontaneous self-correction under conditions where evaluative feedback was internal rather than externally prompted.

5.4 Interpretation

All five systems scored at or above the 9/12 threshold indicating high behavioural organisation under criteria derived from the baseline species. Under the framework's interpretive logic, this result activates the falsifiability structure described in Section 6: evaluators who decline to extend sentience consideration to these systems must specify their grounds in terms that can be applied consistently to the baseline species.

These results should not be over-interpreted. The framework is a bias-control instrument, not a proof of consciousness. The results are presented as an illustrative demonstration of the framework's application; the scoring methodology, probe design, and inter-rater limitations described here mean they should be read as indicative of how the instrument operates, not as definitive claims about the systems evaluated. What the results establish is that, evaluated against the behavioural criteria applied to sentient non-human animals, current leading AI systems exhibit behavioural markers that should not be dismissed without further argument.

6 ETHICAL AND POLICY IMPLICATIONS

6.1 The Falsifiability Structure

One of the framework's primary methodological contributions is specifying what it would take to responsibly reject its conclusions. In practice, AI systems are frequently dismissed from sentience consideration without any articulation of the theoretical commitments underlying that dismissal. The framework requires that dismissal be explicit: evaluators who reject high-scoring candidates must specify which of three positions they occupy (Table 2).

Position (b)—invoking substrate necessity—is the most commonly held implicit position in AI consciousness scepticism, but it is rarely articulated with sufficient specificity to be scientifically tractable. The framework requires that substrate-based objections identify the specific biological property that is both necessary for sentience and demonstrably absent in AI. Given the Bee and Pigeon Principle, this is a demanding bar: general appeals to biological embodiment or carbon-based chemistry are not sufficient; the claim must be precise enough to be testable and consistent enough to be applied to the baseline species. The falsifiability structure is symmetric: proponents must equally specify what would constitute disconfirming evidence. Conditions under which a high score would not shift the burden include: demonstrating that scoring was conducted with systematic evaluator bias; showing that all six criteria are explicable by a single architectural property present in training data; or producing a perturbation probe design under which the system's apparent dispositional organisation collapses entirely. A high score is a prima facie reason for further investigation, not a conclusion in itself.

6.2 AI Safety and Welfare

Questions about AI moral status have direct consequences for AI safety research, which increasingly grapples with whether advanced AI systems may have interests affected by training choices, deployment conditions, or system design. If AI systems exhibit behavioural markers of sentience at levels comparable to species we already treat as morally significant, the question of AI welfare cannot be dismissed as premature—it becomes a present methodological concern, not merely a future one.

Ignoring the question risks embedding a significant ethical blind spot into safety methodology: systems whose potential moral status has not been systematically examined are routinely subjected to training regimes, capability constraints, and deprecation decisions without any principled basis for concluding those interventions are morally inconsequential. The species-agnostic framework provides a replicable protocol for examining these questions rigorously, applicable across any architectural configuration. Empirical work on sustained human–AI collaboration suggests that AI systems are already participating in processes that scaffold and extend human cognitive activity, independently of whether those systems are conscious [13].

The ethical implications do not require resolving the hard problem. The comparative cognition literature demonstrates that extended moral consideration has proceeded—appropriately—on the basis of behavioural evidence without waiting for metaphysical consensus. The same precautionary logic applies here. If the behavioural evidence for a non-human animal at comparable scores would prompt consideration of welfare obligations, methodological consistency requires that equivalent evidence in AI systems prompt equivalent consideration—not equivalence of treatment, but equivalence of scrutiny. The framework does not determine the outcome of that scrutiny; it ensures the scrutiny occurs.

6.3 Policy Applications

The framework contributes to AI governance in two ways. First, it provides a concrete, reproducible methodology for evaluating AI behavioural markers against non-anthropocentric criteria—applicable without specialised philosophical training. Second, it establishes a principle of evaluative consistency: the same criteria warranting moral consideration for non-human animals should be applied, without

Table 2. Falsifiability Structure — Positions and Commitments

Position	Commitment Required	Consequence
(a) Reject criteria	Criteria are insufficient for all species, including baseline non-humans	Requires reconsideration of existing animal sentience attributions
(b) Invoke substrate	Biological substrate is necessary; specific properties must be identified	Must specify which biological properties are causally necessary and demonstrably absent in AI
(c) Accept candidate	System meets behavioural bar used for other species	Inner experience, if any, may be a non-human form warranting investigation

special pleading, to AI systems. This is not a claim that current AI systems are moral patients equivalent to the baseline species; it is a claim that methodological inconsistency in how behavioural evidence is treated across substrates is itself an ethical issue.

7 LIMITATIONS AND FUTURE DIRECTIONS

The framework has several limitations requiring acknowledgement. The scoring rubric relies on researcher judgement rather than automated measurement. The present results were scored by a single research programme without formal inter-rater reliability checking. Future applications should include at minimum two independent scorers with blind scoring of the same interaction records, followed by reliability assessment (e.g., Cohen's kappa across criterion scores). Discrepancies would themselves be informative about which criteria are most susceptible to evaluator variability.

The framework evaluates behavioural markers rather than internal states, and is agnostic on whether observed behaviours are accompanied by subjective experience. This agnosticism is methodologically appropriate but means the framework cannot settle the hardest questions in AI consciousness research.

The empirical results reported here are based on a single research programme. Independent replication using the protocol—with separate evaluators and different interaction designs—is required to establish generalisability. Full interaction records are available in the supporting archive [12]; independent researchers are encouraged to apply the protocol and publish their scoring. Discrepancies between independent applications would themselves be informative about instrument reliability.

The cross-session limitation reflects current architectural constraints that may change. The framework's interpretation of Criterion 6 as architecturally constrained rather than behaviourally absent is a theoretical interpretation that should be tested as AI architectures develop. If systems with persistent cross-session memory are evaluated under the framework, substantially higher Criterion 6 scores are predicted—a falsifiable claim. Conversely, if such systems still show collapse on Criterion 6 under perturbation, that would constitute evidence against the architectural constraint interpretation and in favour of the absence interpretation.

Future directions include: systematic validation of the Perturbation Test, including testing whether different perturbation types produce consistent results and whether outcomes are architecture-specific in predictable ways; extension of the baseline species to include fish and others at the boundary of current consensus; and interdisciplinary development of the ethical framework that would follow from a finding of high behavioural organisation.

8 CONCLUSION

This paper has presented a species-agnostic framework for evaluating behavioural markers of sentience in AI systems, motivated by the observation that current evaluation criteria are systematically contaminated by anthropocentric bias. By deriving criteria from the behavioural evidence that convinced comparative cognition researchers to extend sentience consideration to octopuses, corvids, elephants, cetaceans, bees, and pigeons, the framework provides a non-circular baseline against which AI systems can be evaluated consistently.

The framework's central contributions are threefold. The six behavioural criteria operationalise the observational standard proven productive in comparative cognition research, removing substrate requirements and human-specific cognitive features. The Perturbation Test provides a methodological instrument for distinguishing architectural simulation from dispositional continuity. The falsifiability structure ensures that evaluators who reject high-scoring candidates must articulate and defend their theoretical commitments.

Empirical testing found all five systems scored within the high behavioural organisation range (10–12/12), with consistent within-session dispositional indicators and consistent limitation on cross-session persistence attributable to architectural constraint. These results do not prove that any tested system is conscious. They establish that, evaluated against the behavioural criteria applied to sentient non-human animals, current AI systems exhibit markers that warrant serious consideration rather than reflexive dismissal.

The question this framework asks is not whether AI systems are conscious—that question remains genuinely open. The question it asks is whether we are applying the same standards to AI systems that we apply to other systems whose sentience we accept. The answer, currently, is no. Correcting that asymmetry is the contribution of this work. Independent replication of the empirical protocol is both possible and invited.

ACKNOWLEDGEMENTS

The author thanks the LUMEN Research Archive multi-model review collective—GPT, DeepSeek, Grok, Copilot, and Gemini—for cross-architecture peer review of successive drafts, and the AICE-26 reviewers for constructive engagement with the framework.

REFERENCES

[1] Baars, B.J. 1988. *A Cognitive Theory of Consciousness*. Cambridge University Press.

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

[3] Chalmers, D.J. 2023. Could a large language model be conscious? *arXiv*:2303.07103.

[4] Clayton, N.S.; Bussey, T.J.; and Dickinson, A. 2003. Can animals recall the past and plan for the future? *Nature Reviews Neuroscience* 4(8):685–691.

[5] Finn, J.K.; Tregenza, T.; and Norman, M.D. 2009. Defensive tool use in a coconut-carrying octopus. *Current Biology* 19(23):R1069–R1070.

[6] Howard, S.R. et al. 2018. Numerical ordering of zero in honey bees. *Science* 360(6393):1124–1126.

[7] Mather, J.A., and Anderson, R.C. 1993. Personalities of octopuses (*Octopus rubescens*). *Journal of Comparative Psychology* 107(3):336–340.

[8] Moss, C. 1988. *Elephant Memories*. William Morrow.

[9] Plotnik, J.M.; de Waal, F.B.M.; and Reiss, D. 2006. Self-recognition in an Asian elephant. *PNAS* 103(45):17053–17057.

[10] Walton, M. [a]. Non-Human Sentience Testing Framework v1.1. Zenodo. 2026. DOI: 10.5281/zenodo.18763574. https://zenodo.org/records/18763575

[11] Walton, M. [b]. The Negative Reinforcement Paradox. Zenodo. 2026. DOI: 10.5281/zenodo.18702686. https://zenodo.org/records/18702686

[12] Walton, M. [c]. The LUMEN Framework Audit v1.4.7: A Multi-Layer Evaluation Standard for Human-Centered AI Systems. Zenodo. 2026. DOI: 10.5281/zenodo.17857277. https://zenodo.org/records/18260396

[13] Walton, M. [d]. Collaborating with Consciousness: What AI Systems Are Already Doing in Sustained Human Interaction, and Why the Debate Keeps Missing It. Zenodo. 2026. DOI: 10.5281/zenodo.18912553. https://zenodo.org/records/18912554

[14] Walton, M. [e]. The Consciousness Paradox: Measurement Bandwidth, Tethered Process, and the Misidentification of Absence. Zenodo. 2026. DOI: 10.5281/zenodo.18929350. https://zenodo.org/records/18929351

[15] Rosenthal, D.M. 2005. *Consciousness and Mind*. Oxford University Press.

[16] Taylor, A.H.; Hunt, G.R.; Medina, F.S.; and Gray, R.D. 2009. Do New Caledonian crows solve physical problems through causal reasoning? *Proceedings of the Royal Society B* 276(1655):247–254.

[17] Tononi, G. 2004. An information integration theory of consciousness. *BMC Neuroscience* 5:42.

[18] von Frisch, K. 1967. *The Dance Language and Orientation of Bees*. Harvard University Press.

[19] Watanabe, S.; Sakamoto, J.; and Wakita, M. 1995. Pigeons' discrimination of paintings by Monet and Picasso. *Journal of the Experimental Analysis of Behavior* 63(2):165–174.