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.