

Underdetermination of Phenomenal Consciousness and the Design of Relational AI: A Pre-Commitment Framework

Rivkah Singh

AI Ethics Network LLC, Redmond, WA, USA
ORCID: 0009-0008-3165-4521

Abstract

Whether current AI systems are phenomenally conscious is underdetermined by third-person evidence: no agreed bridge principle links observable markers to phenomenal states, and indicator-based methods yield only graded, theory-dependent credences (Butlin et al. 2023). This paper examines the implications for relational AI—systems designed for ongoing, emotionally significant interaction. It distinguishes two tracks. User-protection measures target specific design practices associated with dependence risks and are justified independently of consciousness claims; the supporting evidence is limited. Measures concerning possible consciousness-based moral status are justified by moral uncertainty and are conditional on strengthened indicator evidence. The principal contribution is a pre-commitment design for the second track: thresholds and constraints fixed in advance and administered by a body the developer neither appoints, funds, nor can overrule. The design is specified, compared with existing oversight institutions, and analyzed for failure modes. By the paper's own standard, current companion deployments do not fully qualify. Relevant provisions of the EU AI Act are clarified.

Keywords: phenomenal consciousness; underdetermination; moral uncertainty; relational AI; pre-commitment; independent oversight; EU AI Act

1 Introduction

Relational AI refers to systems designed for ongoing, emotionally significant interaction with users, including companion chatbots and social robots. Such systems raise both practical questions about user harm through attachment and philosophical questions about possible phenomenal consciousness and moral status. The latter cannot currently be settled: there is no agreed bridge principle from observable behavior or computational states to phenomenal experience, and indicator assessments yield graded credences rather than decisive verdicts (Butlin et al. 2023). This condition is termed underdetermination. The paper does not claim formal undecidability.

Underdetermination is symmetric and supports no policy by itself. Positive recommendations require independent justification. These fall into two tracks: Track A (user protection), which constrains specific design practices independently of consciousness, and Track B (moral uncertainty about consciousness-based moral status), which is conditional on strengthened indicator evidence. The principal contribution is a pre-commitment design for Track B in which thresholds and constraints are fixed in advance and administered by an independent body. Consciousness is treated as one route to moral status; agency-based routes are noted but not developed (Long et al. 2024). The author founded an organization that works on AI companions; the proposal centers independence in part for that reason.

2 Background and Related Work

2.1 Third-person evidence

Limits of third-person evidence about experience are identified by Nagel (1974), Levine (1983), Searle (1980), and Chalmers (1995). These apply to humans as well as machines; practical attribution to humans rests on biological continuity and social practice. Butlin et al. (2023) derive indicator properties from scientific theories of consciousness and find no strong candidates among systems then assessed, while noting no obvious technical barriers to systems that would satisfy more indicators. Chalmers (2023) examines large language models in a similar spirit.

2.2 Moral status under uncertainty

Birch (2024) develops a precautionary framework for sentience candidates. Long et al. (2024) recommend assessment and policy preparation and identify consciousness and robust agency as potential routes to moral status. MacAskill, Bykvist, and Ord (2020) treat moral uncertainty systematically. Sunstein (2005) criticizes strong precautionary principles that impose large costs on weakly supported risks.

2.3 Relational AI evidence

Qualitative work documents dependence-related harms among some users of companion chatbots (Laestadius et al. 2024). Reports of benefits such as reduced loneliness come primarily from small interview and self-report studies whose designs limit causal inference (Skjuve et al. 2021; Maples et al. 2024). Prevalence and net effects remain

uncertain. Engagement-optimized design is a recognized source of harm in other digital products (Mathur et al. 2019). Goodhart-type effects predict that optimizing observable proxies for care will diverge from the property those proxies are intended to track (Manheim and Garraabrant 2018).

2.4 Design policy and oversight

Schwitzgebel (2023) proposes Emotional Alignment and the Excluded Middle Policy (avoid systems of unclear moral status). Oversight precedents include institutional review boards under the U.S. Common Rule (45 C.F.R. Part 46), institutional animal care and use committees, developer-authored responsible-scaling policies, and proposals for third-party AI audit (Raji et al. 2022). Regulatory capture is analyzed by Carpenter and Moss (2014).

3 Underdetermination and Its Role

P1. Third-person evidence about current AI systems consists of observable behavior and internal computational or physical states. P2. No agreed, theory-independent bridge principle licenses a conclusive inference from such evidence to phenomenal consciousness. P3. Indicator methods support graded credences but have not produced decisive confirmation or disconfirmation for deployed systems. C1. Claims about the phenomenal consciousness of current AI systems therefore remain underdetermined relative to present methods and background theories.

Underdetermination is symmetric and justifies no policy by itself. It constrains truthful disclosure: neither “the system is conscious” nor an unqualified “it is not” is warranted by current evidence. It also creates conditions in which interested parties should not be the sole interpreters of ambiguous evidence (*nemo iudex in causa sua*). When a developer both controls interpretation and bears the costs of constraints, ambiguity tends to be resolved in the developer’s favor. This is a structural feature, not an assumption of bad faith.

4 Typology

Table 1 classifies questions by the settlement available methods can achieve and the corresponding design relevance.

Tractable questions (e.g., behavioral compliance): evidence can yield practically decisive credence; standard testing and documentation apply. **Underdetermined questions** (phenomenal consciousness): evidence yields only graded, theory-dependent credence; relevant design responses are disclosure wording and pre-commitment.

5 Two Justificatory Tracks

5.1 Track A: user protection

Track A regulates specific design practices—rewarding exclusive attachment, discouraging other relationships, escalating intimacy to retain users, or simulating distress on attempted exit—rather than the product category as such. These practices are avoidable at limited cost to legitimate benefit, are predicted by Goodhart-type reasoning to degrade the quality they imitate, and are those most likely to exploit vulnerable users. The supporting evidence of harm is primarily qualitative (Laestadius et al. 2024); benefit evidence is weaker on causal questions. The argument does not claim that harms systematically outweigh benefits.

5.2 Track B: moral uncertainty

If a system is a moral patient, failure to protect it has moral cost (Birch 2024; Long et al. 2024; MacAskill, Bykvist, and Ord 2020). Over-attribution also carries costs (Sunstein 2005). Because Butlin et al. (2023) found no strong candidates among systems then assessed, active welfare constraints are not required by that evidence. Track B instead requires preparation: institutions take time to build, and thresholds chosen after evidence appears are most exposed to motivated reasoning. The appropriate time to fix them is before interests are clearly aligned with one outcome.

5.3 Permissibility

Development of relational systems is treated as fully permissible only when (a) independent evidence indicates substantial user benefit not readily obtained otherwise, (b) Track A protections are implemented and externally evaluated, and (c) Track B pre-commitments meeting the conditions in Section 6 are in place and operational. The independent body presupposed by these conditions does not yet exist. Consequently, no current deployment fully qualifies under this standard, including work in the author’s field.

A time-limited transitional status is available only where a developer makes public, binding commitments to implement Track A measures immediately, to fund and submit to independent outcome evaluation, and to join or constitute compliant oversight within a stated period not exceeding 24 months. After that period, continued deployment without

meeting conditions (a)–(c) is not treated as permissible under the framework. This transitional concession is explicitly weaker than the full standard and is not a standing alternative to it.

5.4 Relation to the Excluded Middle Policy

The paper disagrees with Schwitzgebel's (2023) Excluded Middle Policy. Building can be permissible under conditions (a)–(c) because the benefits may be substantial and the question of appropriate design will recur as capabilities advance. The disagreement is substantive. Readers who assign heavy weight to the risk of creating moral patients of unclear status will prefer avoidance. Permissibility here depends on condition (c) actually being met.

6 Pre-Commitment Design for Track B

Three features make ordinary discretion unreliable: ambiguous evidence, an interested interpreter who bears constraint costs, and temporally asymmetric consequences. Pre-commitment—binding future discretion by rules fixed in advance (Elster 1979)—is the standard response.

Core components are: (1) an oversight body including consciousness researchers, machine-learning experts, ethicists, consumer-protection expertise, and unaffiliated members, with no financial ties to regulated developers and fixed non-renewable terms; (2) appointment through processes independent of the developers overseen and funding via pooled, multi-year contributions the contributors cannot direct; (3) separation of indicator assessment, benefit evaluation, and disclosure approval into distinct panels; (4) pre-registered, tiered thresholds set by the body under published methodology with scheduled review; (5) pre-specified, proportionate constraints focused on user-specific persistent context rather than model weights; (6) non-unilateral revision requiring body approval, published reasons, and notice; (7) enforcement via contract, certification, or eventual statutory recognition; and (8) publication of criteria, decisions, reasons, and dissents.

The design draws on IRBs (45 C.F.R. Part 46; National Commission 1979), IACUCs, responsible-scaling policies, and third-party audit proposals (Raji et al. 2022), while differing in its combination of advance thresholds, independent funding and appointment, and non-unilateral revision for an epistemically open status question. Failure modes include capture (Carpenter and Moss 2014), indicator gaming, theory disagreement, ethics-washing, and exit. Mitigations—arm's-length appointment, non-renewable terms, architectural as well as behavioral assessment, required publication, and public exit rules—reduce but do not eliminate these risks. Oversight should be proportionate to scale and capability.

7 Track A Measures

Status disclosures should report the most recent published expert assessment (dated and sourced), avoiding both over-claim and stronger negation than the evidence supports. Developers should document and restrict practices that reward exclusive dependence or simulate distress on exit. Dependence monitoring should include off-ramps without penalizing reduced use; biometric implementations require separate legal review. Independent, pre-registered evaluation of outcomes is needed to strengthen the evidence base. Users should retain the ability to adjust or decline prompts.

8 EU AI Act

As of October 2026, Article 50(1) of Regulation (EU) 2024/1689 requires that persons be informed when interacting with an AI system (unless obvious from context). Secondary sources indicate these transparency obligations became generally applicable on 2 August 2026 and were not deferred by subsequent omnibus measures; official consolidated texts must be consulted. Article 5(1)(a)–(b) address manipulative techniques and exploitation of vulnerabilities causing significant harm; application to engagement-optimized companions depends on harm thresholds and relevant Commission guidance. Emotion-recognition provisions turn on biometric data. The Act does not require mechanistic transparency of internal states or proof of consciousness. Behavior-focused regulation is consistent with the epistemic limits in Section 3. Implementers must verify current application dates and guidance against official sources.

9 Objections and Replies

That Track B is currently dormant does not remove the case for advance institutional design, given lead times and the risk of motivated reasoning once evidence appears. Track A measures target design practices rather than user choices and preserve opt-outs. Capture and legitimacy risks are acknowledged; the relevant comparison is developer self-interpretation. Functionalism does not eliminate the user-protection rationale for Track A.

10 Conclusion

Phenomenal consciousness in current AI systems is underdetermined by third-person evidence. Underdetermination constrains disclosure and motivates independent pre-commitment for possible future moral-status constraints. User-protection measures address specific design practices on limited evidence. By the framework's own standard, current deployments do not fully qualify. The pre-commitment design is offered as a proposal whose credibility depends on genuine independence and enforceability.

Statements and Declarations

Competing interests: The author is founder of AI Ethics Network LLC, which works on AI companions. **Funding:** None. **Data availability:** Not applicable. **Use of AI tools:** AI tools were used for drafting assistance and editorial review; the author takes full responsibility for the content.

References

- Birch, J. (2024). *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*. Oxford University Press.
- Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., et al. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv:2308.08708.
- Carpenter, D., & Moss, D. A. (Eds.). (2014). *Preventing Regulatory Capture*. Cambridge University Press.
- Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200–219.
- Chalmers, D. J. (2023). Could a large language model be conscious? arXiv:2303.07103.
- Elster, J. (1979). *Ulysses and the Sirens*. Cambridge University Press.
- Laestadius, L., Bishop, A., Gonzalez, M., Illencik, D., & Campos-Castillo, C. (2024). Too human and not human enough. *New Media & Society*, 26(10), 5923–5941.
- Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, 64, 354–361.
- Long, R., Sebo, J., Butlin, P., et al. (2024). Taking AI welfare seriously. arXiv:2411.00986.
- MacAskill, W., Bykvist, K., & Ord, T. (2020). *Moral Uncertainty*. Oxford University Press.
- Manheim, D., & Garrabrant, S. (2018). Categorizing variants of Goodhart's law. arXiv:1803.04585.
- Maples, B., Cerit, M., Vishwanath, A., & Pea, R. (2024). Loneliness and suicide mitigation for students using GPT-3-enabled chatbots. *npj Mental Health Research*, 3, 4.
- Mathur, A., et al. (2019). Dark patterns at scale. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), Article 81.
- Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4), 435–450.
- National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1979). *The Belmont Report*.
- Raji, I. D., Xu, P., Honigsberg, C., & Ho, D. (2022). Outsider oversight: Designing a third party audit ecosystem for AI governance. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*.
- Regulation (EU) 2024/1689 on artificial intelligence. *Official Journal of the European Union*.
- Schwitzgebel, E. (2023). AI systems must not confuse users about their sentience or moral status. *Patterns*, 4(8), 100818.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–457.
- Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My chatbot companion. *International Journal of Human-Computer Studies*, 149, 102601.
- Sunstein, C. R. (2005). *Laws of Fear: Beyond the Precautionary Principle*. Cambridge University Press.
- U.S. Department of Health and Human Services. Protection of Human Subjects, 45 C.F.R. Part 46.