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).