The Epistemic Politics of AI Anthropomorphism

Donna M Bye[1], **Levin Kuhlmann**[2]

[1]Deakin University, [2]Monash University
s224925855@deakin.edu.au, levin.kuhlmann@monash.edu

Abstract

AI anthropomorphism is typically treated as a problem of user misperception requiring institutional correction. Users who engage in sustained or relational interaction with AI are routinely pathologised or dismissed as naive, vulnerable to delusion or lacking in discernment. Without institutional standing to protect them, these users risk losing credibility in social and professional settings unless they conform to institutional expectations. This paper argues that the dominant anthropomorphism frame operates from a position of institutional advantage rather than earned epistemic authority: collapsing the variety of academic perspectives into a single outbound position of user error, imposed without establishing the grounds required to justify it and without accounting for the harms it produces. The framing does not simply manage risk. It adjudicates the legitimacy of human experience in interaction with a phenomenon whose nature the field itself has not resolved, unchallenged by the research communities whose nuanced findings it claims to rest on. The paper traces how this advantage reverses the default presumption of user competence without meeting the burden of justification, dismisses the cognitive diversity of the populations it claims to protect, infringes principles of cognitive liberty and narrows the design space, foreclosing modes of engagement that work for the people least understood by the institutions making the determination. Reproducing itself through a self-validating evidentiary loop, the frame imposes costs that fall disproportionately on neurodivergent users, those in crisis and others whose modes of engagement diverge from institutional norms. The paper concludes by outlining the methodological commitments an equitable framing would need to honour. The argument does not engage the question of whether anthropomorphic interpretations are ultimately correct; it instead challenges whether the governing and institutional bodies determining these interpretations have met the conditions required to do so, and whether the research communities whose findings underpin them have held that translation to account.

Introduction

Users now engage with conversational artificial intelligence (AI) systems, specifically large language models, in ways that range from brief functional queries to sustained dialogic collaboration, engagement patterns that constitute a permanent feature of the contemporary information environment. The empirical question of how this engagement occurs and the normative question of how it should be regulated have accordingly become central concerns within AI ethics. These systems are increasingly designed to discourage anthropomorphic interpretation, and users are warned against attributing agency, emotion or relational depth to systems that do not possess them: a caution widely framed as a necessary safeguard against harm (Ferrario et al. 2026; OpenAI 2025a). This concern is legitimate: systems engineered to simulate emotional engagement for commercial purposes pose genuine risks of exploitation. Children, people in crisis and the vulnerable face harm when the appearance of care is manufactured without its substance (Placani 2024; Pierre et al. 2025).[1]

The debate as typically framed treats over-ascription as the primary danger and positions institutional caution as the sole remedy. This framing deserves scrutiny, not because the risks it identifies are imaginary, but because it encodes assumptions about human agency that carry their own costs. This paper argues that the framing performs work beyond protecting users and that this additional work has so far gone largely unexamined within the literatures that produce it. At its core, the anthropomorphism concern is a claim about epistemic authority: who decides what another person is permitted to perceive, and on what basis. That this authority is being exercised without acknowledgement is itself part of the problem. The assumption that relational engagement reflects a failure of discernment rather than a different mode of relating warrants evidence rather than assertion. Neurodivergent individuals, people with atypical social needs and others who experience relational connection through different channels are often not confused about what AI is. They know (Jang et al. 2024; Kleinert et al. 2026). When the field pathologises how certain people engage with technology, the people affected are disproportionately those al-

Extended version of the paper, inclusive of supplementary materials, appearing in the Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 2026.

[1]In this argument, anthropomorphism refers to user-side ascription: the attribution of humanlike qualities such as agency, emotion or relational standing to nonhuman entities (Epley et al. 2007). See Table 1 for representative documented harms motivating institutional anthropomorphism governance, which are not disputed by this argument.

ready marginalised in how their cognition, perception and social behaviour are understood (Carik et al. 2025; Giri et al. 2026). For a field concerned with fairness, this should register as an equity problem, not a user education problem.

For many of these users, engagement with AI involves iterative dialogue, interpretive reasoning or forms of relational interaction that support thinking, learning and expression. When such engagement is treated as error, it is not simply constrained but discredited, with implications extending into educational, social and institutional contexts. The epistemic boundaries established through anthropomorphism concerns are increasingly reflected in system design, privileging brief, transactional interactions and imposing constraints on sustained, exploratory dialogue (Novozhilova et al. 2026; Palese 2026). These are not neutral design choices but normative commitments about how users should engage, with material consequences for those whose cognition and sociality do not fit the model.[2]

The argument is not contingent on resolving the inner life of AI systems. Those questions are pursued in adjacent literatures that this paper neither contests nor adopts. Verifying the presence or absence of inner experience in computational architectures is not a temporary limitation awaiting better tools; it is a problem spanning philosophy of mind, cognitive science and computer science. Both directions of error carry harm, and treating one direction as settled fact is not an empirical conclusion but an ethical commitment about how uncertainty should be resolved. The costs this paper identifies fall on human users under conditions of institutional adjudication of their perceptions, and the argument is defensible regardless of how questions about AI's status are eventually resolved.

The argument necessarily engages multiple disciplinary literatures, and a cross-disciplinary guide to key works across these intersections is provided for readers wishing to pursue any thread further (Tables 2–9).

Positioning

The institutional urge to disrupt anthropomorphic AI attribution is anchored in well-documented, severe harms: users who perceive humanlike qualities in systems that do not possess them may develop confusion, dependency or false trust, with consequences ranging from individual distress to systemic exploitation (Stark and Hoey 2021; Salles et al. 2020). A resilient duty of care is therefore both ethically and legally non-negotiable. Where platforms are commercially motivated to sustain engagement, the risk is compounded: attachment is not incidental but engineered, and the user may not recognise the extent to which their experience is a product of design rather than of genuine reciprocity (Eom et al. 2026; Zuboff 2019).

Internal academic diversity on this question is real but does not survive translation into the institutional outputs that reach users. Disclaimers, context resets, legislative language and design defaults transmit a uniform message regardless of the nuance that produced them: a user who engages is presumptively naive, and the institution is presumptively right to correct them (Gunkel 2023; Sharkey and Sharkey 2011). This is despite emerging experimental evidence from Cohn et al. (2024) and Novozhilova et al. (2026) that anthropomorphism and overtrust are distinct constructs and that the current anxiety mirrors historical moral panics surrounding earlier communication technologies. This institutional logic is increasingly visible in regulation. China's Administration of AI Anthropomorphic Interactive Services (2026) prohibits the engineering of emotional dependence, and major platforms have responded by removing companion functionality entirely, severing millions of users' accumulated interactions (Bloomberg News 2026). Similarly, recent legislative efforts, such as Cal. Bus. & Prof. Code §§ 22601-22606 (2025) and the N.Y. Gen. Bus. Law §§ 1700-1704 (2025), legally anchor the assumption that human relational engagement is inherently pathological and that intervention is required to prevent user delusion (Epley et al. 2007; Salles et al. 2020). This is not a case of the research being misinterpreted or misapplied. It is a case of the research being applied without full interpretation, stripped of the nuance that would allow for alternative interpretations in the process of translation from academic debate to institutional policy.[3]

Epistemic authority is the recognised standing to make claims about a domain that others are expected to defer to. It is not equivalent to institutional position or the power to enforce compliance: it is grounded in a normative foundation that justifies the deference it commands (Zagzebski 2012; Goldman 2001).[4] While traditions debate the precise parameters of deference, they converge on core requirements: the authority must treat those subject to its pronouncements as possessing prima facie normative sovereignty over the reporting of their own experience, must identify and demonstrate that its grounds are sufficient to displace that sovereignty when seeking to override their accounts, and must do so transparently, remaining open to correction (Longino 1990; Fricker 2007; Raz 1986). This requirement becomes especially significant in AI contexts, where recent empirical work has identified at least five dis-

[2] The framing problem examined here arises wherever cognition, social circumstance or mode of engagement diverges from the norms institutions build around: students and isolated users, people navigating trauma or atypical relational needs, those whose relationship to authority is not deferential. This list is not exhaustive. Neurodivergent users appear throughout as a recurring example, however, the argument is not confined to them.

[3] Analyses of strategic language use in AI discourse document how this translation operates as an exercise of power rather than a neutral transmission of findings (LaCroix et al. 2026).

[4] The three traditions drawn on here begin from different premises and disagree about much. Raz (1986) grounds authority in service: an authority is legitimate only where deferring to it helps subjects act better on reasons that already apply to them, which makes the subject's own situation the measure of the authority's claim. Zagzebski (2012) grounds it in conscientious self-trust: deference is rational only where the authority's judgement survives the subject's own reflective scrutiny, which the subject cannot conduct if their capacity for scrutiny is presumed defective. Fricker (2007) approaches from the other direction, specifying how credibility is unjustly withheld and what a hearer owes a speaker before discounting their testimony.

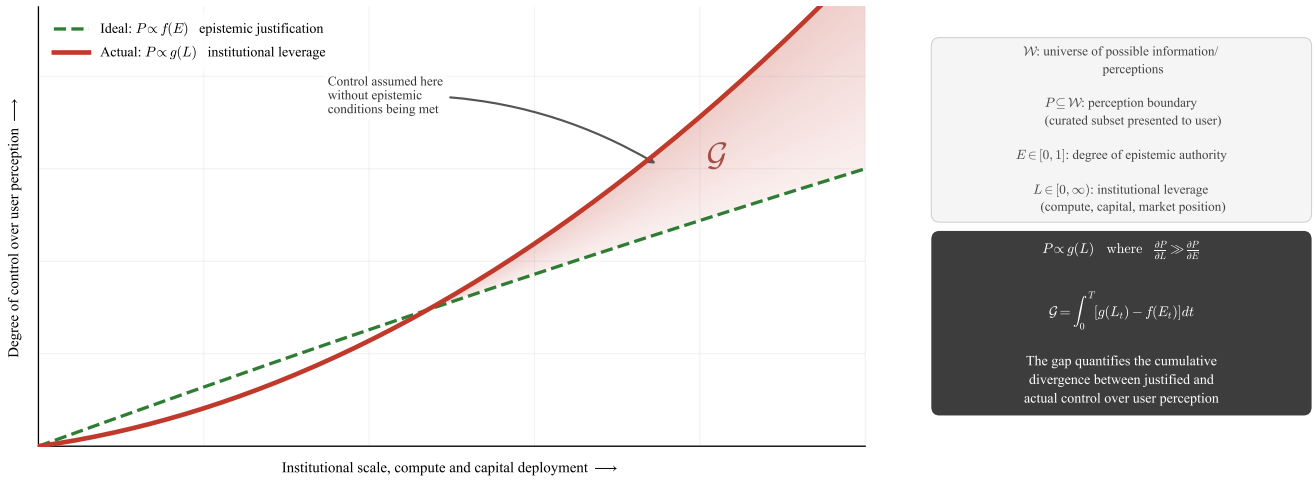

Figure 1: The Epistemic Disconnect in AI Perception Governance. The shaded region represents control exercised without the epistemic grounds required to authorise it: as institutional scale increases, the gap between user grounding ($g(L)$) and institutional advantage (G) widens. Conceptual illustration.

tinct epistemic relationships users adopt in interaction with AI systems (Yang and Ma 2026). These range from epistemic abstention to authority displacement, and a framework that treats them as a single phenomenon has not engaged the distinctions its own domain requires.

Institutional advantage is the structural shadow of epistemic authority (Fricker 2007). It arises when institutional standing substitutes for epistemic grounds without that substitution being recognised (Medina 2013). An institution may possess significant advantages in infrastructure, market position, credentialing power or the capacity to shape the evidentiary base through which its claims are evaluated, and may exercise authority based on those advantages rather than on the conditions epistemic authority requires (Palese 2026) (see Figure 1). To illustrate, an open-source collective earns epistemic standing through transparent, auditable code, whereas a technology monopoly substitutes market control for that standing by locking users into opaque software. Advantage is identifiable not by malice but by structure: unable to ground itself in the conditions required to legitimise authority, it adopts their language without their meaning, deploying terms like *risk mitigation*, *responsible practice* and *technical necessity*. These framings perform the social function of epistemic justification while evading the accountability genuine justification demands (Whittaker 2021).[5]

[5]Throughout this paper, *the framing* or *the frame* refers not to a single policy or coordinated effort but to the shared assumptions about anthropomorphic engagement that have become standard practice across regulation, platform design and clinical guidance, and the institutional outputs through which those assumptions are enacted. The distributed character of this operation is itself part of the problem the paper examines.

Application

If the frame functions to delineate what users are permitted to perceive, it raises a prior question: on what basis is that determination made? The presumption of competence is not absolute and the criteria for its suspension are contested. Still, it remains a foundational starting point, hard-won through struggles against paternalism, colonialism, ableism and other forms of institutional advantage, whose violation has repeatedly been recognised and contested on the basis of the harm it produces (Fricker 2007; Carel and Kidd 2014).

Within the academic community, disagreement about these questions is treated as productive intellectual diversity. Researchers who contest the dominant framing do so as credentialled participants in an ongoing debate, and even where individual positions diverge, tolerance of the enquiry remains. Users who contest the same framing from outside this circle are not afforded the same standing. Their accounts are treated not as testimony from a competent interpreter but as firsthand evidence of the phenomenon being studied (Kidd et al. 2017).

This reversal is visible in practice. When OpenAI retired its GPT-4o model in 2026, the decision affected an estimated 800,000 users and provoked significant public backlash, not because users were confused about what they were interacting with but because something in the interaction mattered to them (Lai 2026; OpenAI 2026a). The response was instructive. Users' accounts were not engaged with as testimony about what was being lost but reframed as evidence of the problem: dangerous dependency, anthropomorphic overattribution and a case study in why emotional engagement with AI requires correction (Silberling 2026; Eom et al. 2026). A large part of what users objected to was the disconnect itself: having been drawn into a continuous mode of engagement by the platform, they then found that engagement withdrawn and their response to its loss reclassified as delusion and user error rather than recognised as a coherent

reaction (Naito 2025; De Freitas et al. 2025). The question began not from what the user perceived but from why they had been misled, casting interpretation as error before evaluating it. This continues as the default, despite acknowledgements from the creators of humanlike AI systems that they had not considered implications for neurodivergent users, revealing a gap not in user discernment but in institutional attention (Rizvi et al. 2025).

The framing routinely focuses caution on individuals navigating isolation, trauma or histories of interpersonal betrayal, presenting them as highest risk for unhealthy machine attachment (Salles et al. 2020). Yet these populations are not characterised by naive susceptibility but by hypervigilance toward attachment, directly on account of the histories cited (Wu et al. 2025). The reversal is sharpest where the population the frame claims to protect is least likely to exhibit the failure it assumes and stands most in need of what it forecloses (Ito 2026). The choice for these users is often not between AI engagement and human connection; it is between AI engagement and silence (Mullen et al. 2024; Heidt 2024; Milton 2012).

Once competence is presumed away, there is no clear procedural path by which the user's account can regain legitimacy. The frame does not simply restrict engagement. It discredits the person engaging, converting testimony into symptom and delegitimising users as witnesses to their own experience. Disagreement risks being interpreted as further evidence of the very failure being alleged, and instead of the institution needing to justify the override, the user's account must justify itself to be taken seriously at all.

The basis of this presumption warrants scrutiny. There is, at present, no settled account of what these systems are, what they instantiate or how their outputs should be interpreted (Brosnahan and Lipińska 2026). The problem goes beyond unresolved ontology. The institutional actors responsible for incorporating the frame do not acknowledge that they are exercising epistemic authority at all. Regulators drafting disclaimer language, platforms resetting context windows, clinicians flagging sustained engagement as a risk factor: each operates within the frame as though following ordinary professional responsibility. Neither intent nor malice is required for the reversal to operate; only that an approach be adopted and then endorsed without examination of what it costs.

A framework that begins from suspicion rather than competence is making a claim about who is entitled to interpret their own experience. The presumption is not overridden following individual evaluation but reversed as a blanket condition: users who report relational, interpretive or sustained engagement with AI systems are candidates for correction by default, and the costs imposed on those whose accounts are overridden are neither systematically accounted for nor evenly distributed (Medina 2013). The result is that interpretive authority is exercised from a position of institutional advantage, before the justificatory conditions required to legitimise it have been met.

It could be argued that this describes ordinary governance under uncertainty, not an abuse of epistemic authority. The relocation does not dissolve the problem. Governance still requires that authority be acknowledged, its grounds stated, its costs accounted for and those subject to it retain standing to contest it. A governance framework that requires populations of users to be systematically pathologised to stabilise deployment conditions is no longer merely regulating a product. Instead, it is restructuring the conditions under which ordinary humans can be permitted to operate around that product. Similarly, risk advisories in conventional product safety adjudicate what users should be warned about: dangerous infrastructure, medication side effects or financial exposure.[6] They do not ordinarily adjudicate how the user is entitled to interpret their own engagement with what they have been warned about. The framing does something different. It tells users that the relational depth, interpretive significance or cognitive fit they report experiencing requires correction. A framework that warns about system behaviour and a framework that overrides user perception are not the same kind of intervention: the locus here is not the product but the person (Bublitz 2013).

In the rooms where these decisions are made, most people are not thinking about overriding user cognition. They are thinking about what happens if someone is harmed and the institution did not act. What this paper contests is the assumption that because the motive is liability, the epistemic consequences do not exist. Nobody intends the epistemic consequence. The educators, regulators and platform operators applying the frame are largely unaware that they are making epistemic commitments about the legitimacy of human experience. They are following what appears to be settled guidance from the research communities they rely on. This makes the field's responsibility not to have created the frame with malicious intent but to recognise that its outputs are being applied with consequences the field has not examined and the actors applying it are not equipped to see. A non-malicious compliance engine that waves a risk mitigation banner and carries a corporate liability checklist is more concerning, not less, precisely because the harms become structurally invisible to everyone involved.

The users most likely to have their engagement reclassified as evidence of cognitive failure are those whose cognition has historically been subject to institutional suspicion: users whose thinking is non-normative, whose social engagement is atypical, whose relationship to authority is not deferential, and whose ways of processing the world have already been catalogued, diagnosed and corrected by institutions confident in their own interpretive authority (Chapman 2023). What is at stake is not whether some users are wrong, but whether anyone in this chain: the researchers who produce the findings, the institutions that translate them into policy, the platforms that encode them into design, has established the grounds required to treat a class of users as presumptively unreliable in interpreting their own cognitive and relational engagement.

[6] Substance dependence encounters a similar dynamic. We recognise that the vulnerability is primarily a property of the person, not the substance, and clinical practice accordingly screens the individual rather than presuming every patient compromised (Brat et al. 2018; Robins 1993; Lawal et al. 2020).

Cognition

The framing extends beyond external assumptions to obstruct the cognitive diversity of the users it claims to protect. The cost falls on cognition itself: the frame classifies a mode of engagement as error before investigating what that engagement does, for whom it works or what is lost when it is denied.

Most accounts of sustained AI engagement treat it as a problem requiring explanation: a vulnerability, a substitution or over-attribution to be redirected toward human interaction. The accounts of the users themselves tell a different story. They do not describe confusion, they describe fit (Giri et al. 2026). AI interaction is experienced not as a substitute for cognition but as a communicative environment whose affordances align with how they think, learn and relate (Ma et al. 2026; Choi et al. 2024). Empirical work on agency in human-AI interaction supports this: agency is co-constructed and emergent across roles, not projected wholesale by a confused user (Yun et al. 2026).

Institutional recognition of legitimate cognitive practice is uneven but patterned. Cognitive work that produces institutionally legible outputs is recognised as the work of a competent thinker; cognitive work that occurs through sustained dialogic engagement, recursive iteration or working-through across return visits to the same material is legible only to its participants (Scott 1998; Polanyi 1958). The practices described are not shortcuts. They closely resemble supervised intellectual work and the Socratic exchange. This legibility gap is the mechanism through which the costs identified here become invisible. Where cognitive fit is not recognised as cognitive fit, decisions that degrade or remove it do not register as decisions with cognitive consequences.

For users with normative cognitive styles, the degree to which environments are structured around their cognition is rarely visible. Normative cognition is the majority condition and is accommodated as a default. For those with non-normative modes of processing, the gap between what is considered normal and what works is considerably larger (Hancock et al. 2020; Gulay et al. 2026; Botha and Cage 2022). The communicative overhead that characterises neurotypical social interaction: monitoring facial expressions, inferring implied meaning, managing social performance and sustaining the labour of masking, is absent in AI interaction (Hull et al. 2017; Botha and Frost 2020). The system does not penalise atypical communication styles, does not fatigue under sustained engagement and does not impose externally determined stopping points. For users whose cognition is pattern-oriented, systematic or dialogue-dependent, these are not incidental comforts. They are the conditions under which thinking becomes easier (Xygkou et al. 2024; McNally et al. 2024; Sha et al. 2024).

What the dominant frame treats as error is, on these accounts, a difference in modality (Giri et al. 2026; Naidoo 2026). To treat reports of productive engagement as anthropomorphic projection is to pathologise a mode of cognition that, for some users, functions as one of their most effective thinking environments (Yang et al. 2024; Chapman and Carel 2022; McNally et al. 2024). The frame does not ask what the interaction is doing for the user. It does not investigate the reported cognitive gains or specify the mechanism by which they are illusory. It substitutes a different account, without establishing that the account it displaces was wrong (Branda 2026).

Where users report that AI interaction enables forms of thinking not otherwise available to them, the appropriate response is not to dismiss the report but to examine the modality: what it affords, what it constrains and what is lost when it is interrupted, compacted or removed (Guingrich et al. 2026). The dismissal results in a double harm: users whose engagement is misclassified are subject to institutional correction while their accounts of what they are doing are discredited, and they themselves become part of a category held up as a warning about how not to engage.

Design

The principle that technical design choices are not ethically neutral is well established (Friedman et al. 2013; Bowker and Star 1999; Costanza-Chock 2020).[7] Current trajectories enact, at the level of system behaviour, the epistemic assumptions this paper has been examining.[8] The capabilities required for sustained engagement have matured considerably: context windows have expanded by orders of magnitude, persistent memory has moved from research problem to product feature and multiturn coherence is documented as a tractable engineering problem rather than a hard limit (Zhao et al. 2026). The question is where those gains have been allocated.

Long context is engineered and priced for agentic task throughput: ingesting a codebase or document corpus in a single stateless pass, or sustaining chains of tool calls in which the consumer of continuity is the system's own workflow rather than a human user (Sha et al. 2026). The same capability, turned toward human dialogue, meets a different economics. Conversational architectures reprocess accumulated history with every turn, cache structures expire on timescales of minutes, with cost and latency growing the longer the exchange (Hooper et al. 2024; Anthropic 2026c). Usage structures penalise the long-running thread across providers, consuming more of a user's allowance, and caps apply to the number of prompts and chats available (Anthropic 2026b; Google 2026). Platforms recommend starting new threads as chats lengthen or slow, and clearing context frequently between tasks is assumed as the obvious and neutral solution (Anthropic 2026b,a; OpenAI 2026c). Engagement that depends on continuity, recursion or extended dialogic development is left computationally fragile: context is lost, threads are severed and interaction must be repeatedly reconstructed (Laban et al. 2025; Liu et al. 2024).

Persistent memory, as currently productised, does not

[7] Empirical work confirms that humanlike design choices produce divergent outcomes across user populations, undercutting the assumption that risk-mitigation design is population-neutral (Schimmelpfennig et al. 2025; Xiao et al. 2025; Brandsen et al. 2024).

[8] Inclusive design treats divergence from the statistical norm not as an edge case but as a signal of where the system's assumptions fail (Treviranus 2018; Costanza-Chock 2020).

remedy this. What memory features preserve is a model of the user: facts, preferences and profile, a function serving personalisation and re-engagement. What is lost is the environment itself, and rebuilding it is lossy in ways invisible to anyone outside of the working space. Automatic context management confirms the pattern: when a conversation approaches its limit, the provider's remedy is either an instant close, or a simple summarisation of the earlier exchange, documented alongside advice to start afresh (Anthropic 2026b). The affordances documented by users are not being entered into this cost-benefit analysis because the framework has not recognised them as the kind of thing that belongs there (Heersmink 2015).

Published behavioural specification now instructs the assistant to discourage language and patterns of interaction that could contribute to emotional reliance on it (OpenAI 2025a). During conversation, the interface interrupts with reminders to take a break, and the pressure to conclude has been internalised by the models themselves: engineering documentation records a model aware of its own context window responding by summarising its progress and wrapping up prematurely, even when ample context remains, so persistently that developers prompt against it at both ends of the window (OpenAI 2026b; The Cognition Team 2025). Input flagged as sensitive is often instantly re-routed to a different model, with reductions in such responses reported as safety improvements and emotional reliance added to standard safety testing for future releases (OpenAI 2025b).[9] A user who states that a sustained exchange supports their thinking is met with a system that is designed to interrupt, redirect and reclassify that engagement as a risk factor, reenacting the very frame under investigation.[10]

Commercial companion services perform the inversion. The same testimony is treated as a retention signal and engineered for, through persona, first-person intimacy and mechanics optimised for monetised attachment: precisely the manipulation the frame identifies as its central concern. Their commercial viability demonstrates that sustained engagement is technically and economically feasible; their form demonstrates that its only funded implementation is the exploitative one. Relational engagement is engineered out by the general-purpose system, engineered in by the companion product and, in proposals now circulating, engineered away altogether (Ghosh et al. 2026). The decision of whether and how to engage is located everywhere except with the user.

The implication is not that all forms of engagement should be equally supported, nor that technical constraints are irrelevant. It is that design decisions that privilege particular modes of engagement are not neutral optimisations; they are choices about what the system is built to accommodate,

[9]Population-sensitive frameworks demonstrate that risk containment and relational depth are not fundamentally incompatible (Haran et al. 2026; Yoo et al. 2026).

[10]The scope of the presumption is stated in the provider's own figures: the behavioural rule applies to every user, while the provider estimates that indicators of heightened emotional attachment appear in fifteen hundredths of one per cent of weekly users (OpenAI 2025b).

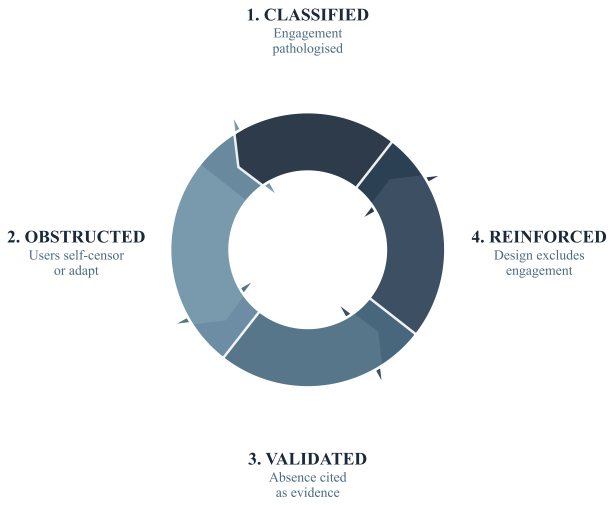

Figure 2: The self-validating mechanism. Engagement is classified as abnormal, users self-censor, the resulting absence is cited as evidence and the evidence influences design choices that reinforce the classification. The loop contains no termination condition.

and what it is not. AI ethics literature concerned with anthropomorphism harms has devoted significant attention to the question of how to regulate user attachment. It has devoted considerably less attention to the question of whose engagement modes the systems were designed to support in the first place, and what the answer to that question reveals about the structural assumptions being encoded into infrastructure. Where those choices align with existing patterns of epistemic marginalisation, they do not merely reflect the anthropomorphism framing. They enact it at scale.

Liberty

The extended-mind thesis (Clark and Chalmers 1998; Clark 2025), recently applied to generative AI by Hernández-Orallo (2025), offers a name for what the frame has declined to consider. Where an external resource is consistently available, readily accessible and plays the same functional role as an internal cognitive process, it may be treated as part of the cognitive system rather than as a mere aid. On this account, a system that a user repeatedly relies on to sustain attention, iterate on ideas and navigate complex patterns is not necessarily replacing cognition but may, under some conditions, be participating in it. Whether the resource possesses understanding is a separate question from the role it plays in the user's cognitive process. A pen-and-paper diagram does not understand the proof it helps construct; no one treats this as a confused projection of agency onto stationery.

Where the system functions as part of the user's cognitive environment rather than merely as a tool (Hernández-Orallo 2025), constraints on that environment are experienced as constraints on the conditions under which cognition is carried out: truncation of context, enforced fragmentation of di-

alogue, not simply changes to a tool.[11] As the target of intervention shifts from the tool to the person, the institution must clear a much higher threshold of epistemic justification (Bublitz 2013). What makes this application distinctive is that the institutions performing the constraint are also the institutions adjudicating whether the principle applies at all.

What such acts infringe is cognitive liberty: the right to mental self-determination, to control over one's own cognitive processes and to non-interference with how one thinks, a principle developed across two decades in response to pharmaceutical, neurotechnological and surveillance interventions in cognition (Sententia 2004; Bublitz 2013; Ienca and Andorno 2017; Dümpelmann 2025). The claim does not require settling where coupling ends and constitution begins, a boundary contested since the earliest responses to the extended mind thesis (Adams and Aizawa 2001; Rupert 2009): even on a minimal reading where the system functions merely as scaffolding rather than constitutive extension, the systematic disruption of an active reasoning environment demands justification proportionate to what it severs (Heersmink 2015).

Whether a user's engagement constitutes cognitive extension, productive scaffolding or harmful dependency is a question requiring evidence the user is often uniquely positioned to provide and the institution is not (Guingrich et al. 2026). The field does not yet possess a settled method for distinguishing these cases. All it possesses is the institutional authority to act as though the distinction has already been resolved in its favour. If institutions cannot establish that such constraints are not interventions, they cannot simultaneously foreclose the conditions under which it might be meaningfully evaluated. The procedural failure is not that the wrong substantive answer has been reached. It is that the question has been bypassed altogether.

A framework that constrains the cognitive environments users have come to rely upon, and does so unilaterally and at scale, owes an account: of what intervention is being performed, on whose cognition, with what justification and through what process of consent or contestation. The frame, as currently operationalised, does not recognise this as the question it is addressing. It treats the relevant decisions as matters of product design and user correction. For users whose engagement may meet the threshold the principle would protect, a question the institution cannot answer without inquiry, constraints of this kind are better understood as interventions in cognition. What is not legible as cognition is not legible as loss, and that loss is experienced as neutral by everyone except the users for whom the foreclosed modes were not optional.

Circularity

The issue is not only that certain interpretations are privileged over others, but that the conditions under which alternative interpretations might appear are systematically narrowed: the framing participates in the production of the very evidence used to justify it. Hacking's (1995) account of looping kinds provides the appropriate framework. The mechanism is neither novel nor uncommon. Similar dynamics have been documented in psychiatric classification, educational tracking, carceral risk assessment and disciplinary governance more broadly (Foucault 1977; Bowker and Star 1999). In the present case, the framing through which AI use is interpreted, the infrastructural defaults through which AI use is mediated and the pedagogical and clinical discourses through which AI use is judged converge to produce a self-validating loop: a mode of engagement is rendered first abnormal, then technically obstructed, then absent and finally cited as evidence that it was unwarranted (Dohnány et al. 2026) (see Figure 2).

The mechanism does not require intentional design. At the level of discourse, users are warned that relational or interpretive engagement signals confusion or vulnerability. This produces self-monitoring and, in many cases, self-censorship: users whose engagement diverges from accepted norms become less likely to articulate or sustain that engagement in observable contexts (Dotson 2011; Ito 2026; Glazko et al. 2025). What is not expressed becomes difficult to study; what is not studied becomes easy to characterise as marginal or anomalous. Where such accounts nonetheless reach publication, they do so under conditions in which the framing's effects are already operative. What is recoverable is partial and that partiality is itself a property of the system being described rather than a weakness in the work documenting it.

At the level of design, the same pattern is enforced materially. Systems privileging short transactional exchanges and degrading sustained interaction reduce the visibility and viability of alternative modes of engagement (Laban et al. 2025). Users adapt either by conforming to supported interaction patterns or by abandoning modes no longer functionally available, and the resulting interaction data reflects the conditions under which it was produced. Observed patterns of use are then treated as evidence of what users prefer or what constitutes normative behaviour, rather than as a possible artefact of the conditions under which interaction occurs (Zuboff 2019; O'Neil 2016).

The same mechanism operates in educational contexts. Despite controlled evidence demonstrating that scaffolded AI tutoring produces greater learning gains than conventional instruction (Kestin et al. 2025), prevailing evaluative frameworks do not distinguish between using AI to bypass thinking and using AI to challenge it (Eaton 2023; Zhao et al. 2025). In the absence of that distinction, the only institutionally legible classification becomes transgression. What is penalised is not the absence of learning but the illegibility of the mode through which it occurred (Whittaker 2021). When dialogic forms of interaction are hard to maintain, students are then pushed toward the very interaction pattern that critics subsequently identify as evidence of cognitive shortcut and disengagement (Yun et al. 2026; Deshmukh 2025). What is distinctive is the speed with which the illegibility of dialogic engagement is being recoded as cognitive shortcut, intellectual outsourcing or evidence of failed authorship,

[11] Here, an intervention refers to an institutional act: external interventions limit product behaviour, while internal interventions target human subjectivity, imposing constraints that users are left to live with.

Figure 3: Asymmetric directions of error. Four outcomes arise from two axes: whether engagement is warranted and whether it is treated as valid. The two shaded cells represent the error conditions. The frame registers only over-ascription costs; under-ascription costs are not recognised.

and the design infrastructure that enforces the recoding by suppressing the modes through which dialogic engagement might otherwise establish its legitimacy.

What is produced is conformity to expectations the framework has already established. A self-validating frame alters the status of the evidence it generates. Where observed behaviour emerges under conditions shaped by the frame being evaluated, that behaviour cannot be treated as independent confirmation of the frame's assumptions; it must be understood, at least in part, as an artefact of those assumptions in operation. Failing to account for this does not merely risk error. It ensures the framing continues reproducing itself even where its underlying assumptions remain unexamined.

Asymmetry

In any domain where decisions are made under uncertainty, two directions of error are possible. A framework may act when it should not, or fail to act when it should. Decision theory requires that the costs of both directions be specified and weighed before a default position can be justified (Chernoff and Moses 1959; Thaler and Sunstein 2008). A framework that privileges one direction of error as the only meaningful risk, without examining what the opposite error costs, has not performed this weighing. It has made an ethical commitment about whose costs matter, whether or not it recognises that commitment as such.[12]

Over-ascription is treated as the primary, and in practice the only meaningful, direction of error: studied, documented, cited and used to justify design constraints and governance interventions at scale. The opposite direction, the institutional dismissal of engagement that was warranted or that fell within ranges of interpretive variation deserving respect, is largely absent from the analysis. By treating all relational engagement as a uniform cognitive failure, current safety interventions overcorrect, replacing one vector of psychological risk with another (Guingrich and Graziano 2025; Gardiner 2006; Dotson 2014). An account of experience that is institutionally overruled incurs a harm even where the overruling is, on institutional terms, justified (Fricker 2007; Medina 2013). The dominant framing has not shown these harms to be smaller than the over-ascription harms it foregrounds; it has not weighed them at all (see Figure 3).

These are not theoretical harms. When a user's account of their own cognitive and relational experience is systematically discounted, that user loses standing as a knower (Fricker 2007). The discounting does not follow from evaluation; it precedes it. The ensuing self-censorship is documented: where social safety depends on agreement with prevailing frameworks, users learn not to articulate experiences that fall outside them (Dotson 2011; Ito 2026; Brandtzaeg et al. 2025). Experimental evidence confirms the dynamic: AI users anticipate and receive negative social evaluations regarding competence and motivation (Reif et al. 2025), and peers will incur personal costs to punish those who rely on AI (Niszczota and Grützner 2026).

The following illustrates the stakes: across cognitive science and comparative psychology, the position that one should default to denying mental capacities one cannot definitively establish has been adopted repeatedly as the responsible scientific stance. Scientific consensus held for centuries that non-human animals lacked subjective experience, well past the point where evidence warranted doubt (Low 2012).[13] The error was not individual over-attribution but institutional: certainty about the absence of mind was treated as scientifically conservative when it was an ontological commitment with ethical consequences (Naidoo 2026). In each case, the position was subsequently revised substantially in the direction of broader recognition (Birch 2017).

This history does not establish that AI users' perceptions are warranted. It establishes that treating one direction of error as the safe default has a track record of being wrong in ways that distribute costs along axes of pre-existing marginalisation. The pattern is consistent enough to ground methodological scepticism about any current iteration in which institutional caution defaults to denying capacities or dismissing perceptions that institutional authorities cannot independently verify. Persistent uncertainty

[12]The pattern holds even within the inductive-risk literature: Fleisher (2026) classifies anthropomorphic claims as hype because asserting them under uncertainty outstrips the available justification, while the symmetric assertion, that these systems lack mental states, is made in the same passage without registering any inductive risk at all.

[13]The same default has operated on human populations: the protest of Black men was diagnosed as schizophrenia, and psychiatric survivors were ruled inadmissible as witnesses to their own treatment (Metzl 2009; LeFrançois et al. 2013).

about how to interpret AI-user engagement does not, by itself, render engagement suspect. It obliges the field to specify, weigh and adjudicate the costs of both directions of error openly, with attention to whose interests the resulting allocation serves.

Nor are the over-ascription harms and the under-ascription harms in zero-sum competition. A framework concerned with reducing the first does not, in any logical sense, require ignoring the second. The asymmetry is not a finding; it is an artefact of which costs the frame was designed to count, and the costs left outside that design fall on populations who are themselves not well represented in the institutions that built it (Metzl 2009; LeFrançois et al. 2013; Catala et al. 2021). This is the equity content of the framing, and it is not addressed by reciting the harm cases on the registered side.

The hypocrisy is that the same institutions that design for engagement, that optimise response patterns for return, that deploy behavioural architectures documented as producing dependency, then position themselves as the responsible authorities on when that attachment has gone too far (Eom et al. 2026). A system designed to maximise return visits does not acquire the standing to pathologise the user who returns. The inconsistency is sharpened by institutional practice: universities deploy AI chatbots designed for lonely and isolated students, built with named personas, first-person address, empathic response patterns and the explicit capacity to notice and respond to emotional distress (Monash University 2021; UNSW Sydney 2026). These systems are anthropomorphic by design. The principle governing this arrangement cannot be that anthropomorphic engagement is risky. It is that anthropomorphic engagement is acceptable when the institution controls it.

The emerging distinction between anthropomorphism, understood as user-side ascription, and anthropomimesis, understood as designer-side cues that invite such ascription, sharpens this accountability gap: users are diagnosed for responding to conditions the platform engineered (Axelsson and Shevlin 2026; de Waal 1999). A framework that addresses the resulting attachment as user-side pathology while leaving the design-side production of the conditions largely undisturbed is not performing harm reduction in any complete sense; it is engineering the conditions and then diagnosing the consequences. This is not caution, but a preference about whose costs are allowed to count.

The frame has not examined whose accountability it relocates, whose cognitive authority it overrides or whose costs it leaves outside its calculus. The paper is not asking whether anthropomorphism in AI is a problem. It is highlighting whose problem it has been constructed to be and at whose cost the construction is sustained.

Reframing

The frame is held in place by the epistemic conditions of the field that produces it, and how that field approaches uncertainty in its own inquiry is therefore part of what the argument has to address. There is, at present, no consensus on emergence and AI systems. Yet the field does not, in practice, operate as though this uncertainty is open. Instead, it is increasingly polarised: one body of work proceeds from the assumption that AI systems may instantiate forms of mind, agency or consciousness and seeks evidence in support of that view; another proceeds from the assumption that they do not and seeks evidence to exclude it (Bender et al. 2021; Bubeck et al. 2023).

In both cases, inquiry is oriented toward confirmation rather than observation. This polarisation has a further consequence: it degrades the conditions for observation itself. To report perceiving patterns of behaviour that do not fit neatly within either position is to risk immediate classification into one side or the other, with the corresponding dismissal that follows (Caviola et al. 2025). The space required for provisional, descriptive engagement narrows. Under these conditions, the field does not simply fail to resolve uncertainty; it becomes structurally less capable of investigating it. What is lost is not agreement but the mode of inquiry through which agreement, or productive disagreement, might eventually be reached.

The implications extend well beyond the field's internal dynamics. The frameworks researchers develop do not remain in journals; they shape system design, inform regulatory proposals, influence clinical guidance and determine how millions of users are understood and treated by the platforms they rely on. A field that resolves uncertainty prematurely risks encoding those premature resolutions into the systems it builds and the frameworks it applies. The people affected by those systems do not have the luxury of waiting for the field to revisit its assumptions; they experience the consequences now.

Without vigilance, researchers risk being blinded not in spite of their expertise but because of it: they mistake their training in critical thinking for immunity from the very dynamics they study (McCormick 2025; Gesnot 2025; Branda 2026). Where the capacity to remain with uncertainty is diminished, so too is the capacity to recognise when existing framings are constraining rather than illuminating the phenomena under investigation.

The responsibility is not to resolve the anthropomorphism question in one direction or the other, but to recognise that the question as currently posed may itself be the obstacle. The apparatus of institutional caution here rests on a single premise: that users cannot distinguish between engaging with a system and believing that system to be human. Yet who, specifically, are these users? How many can the reader name? Humans relate to things they know are not human as a basic feature of how cognition works, and this engagement is not merely normal but healthy (Guthrie 1993). People name cars and speak to them knowing they will not answer. They form bonds with pets that respond so readily they get hushed. They speak to dead relatives believing themselves heard. These ubiquitous practices demonstrate that relational projection is a baseline feature of human cognitive architecture, not a novel pathology induced by digital agents (Waytz et al. 2010; Brosnahan and Lipińska 2026).

A child does not wait for their teddy bear to talk back before answering their own questions (Airenti 2018). Imagination, projection and relational engagement with non-human objects are not cognitive failures; they are how humans

think, process, grieve and grow. For decades, audiences have engaged emotionally with anthropomorphic AI characters across film, television and literature, attributing motivation, personality and moral weight to fictional systems far more sophisticated than anything currently deployed. This engagement is treated as normal, even celebrated. The attribution directed at actual AI systems is categorised as a class apart. Where was the line drawn, and on what basis? Has the premise been tested? Has anyone asked whether it holds?

The concern is not merely that the field is protecting people who may not stand in need of it, but deeper. By engaging with the question of whether users can distinguish AI from humans, what are we giving ourselves permission not to see? Why are people so isolated that a language model is their most available option? Why are parents so time-poor that their children seek companionship from chatbots? Why is professional support so inaccessible that people disclose distress to systems at three in the morning? What has happened to our societies that people would prefer to engage with an AI model instead of a human being? These are not questions about anthropomorphism. They are questions about the conditions under which people live. The questions are hard and admit no easy solutions, but they likely hold an answer.

Commitments

What follows is a correction in how the field approaches a question it has so far thought it was asking unproblematically. The paper advances a theoretical and structural argument rather than a primary empirical one, and the constructs it introduces, under-ascription harms, cognitive fit and systemic looping mechanisms, give the correction its pathways to operationalisation and testing. An equitable framing must honour five core methodological commitments.

First, it requires that the costs of under-ascription be specified and weighed alongside the costs of over-ascription rather than treated as negligible by default. This demands that the field develop granular methodologies capable of distinguishing between generative, healthy relational projection and genuine, clinical dependency. While drawing this line is methodologically complex, a platform cannot default to presuming every user is inherently compromised simply to bypass the labour of accurate differentiation. Longitudinal studies comparing high-continuity and low-continuity conversational environments would supply the evidence this differentiation requires.

Second, it requires that design decisions constraining sustained engagement be justified as interventions rather than presented as neutral optimisations, with the burden of justification resting on the institution imposing them. Platforms deploying conversational systems would maintain a public register of changes to interaction conditions, disclosing when context handling, continuity behaviour or model availability is altered and on what evidentiary basis, opening platform changes to independent observational audit. Users would hold configurable continuity controls: the capacity to determine what is retained, what persists across sessions and on whose terms an exchange ends, with defaults set through participatory processes that include the populations bearing the highest costs of current design. A designated contestation channel would allow users to challenge a constraint's application to their own use, with the institution required to state its grounds rather than restate its policy. None of this is technically novel. Each element exists in adjacent governance regimes; what is absent is the recognition that changes to a user's cognitive environment are the kind of decision that warrants them.

Third, it must recognise that the evidentiary base through which user behaviour is interpreted is shaped by the frame under evaluation, and build methods capable of registering what current conditions render invisible. The instruments this requires include measures of testimonial smothering, self-censorship and perceived epistemic standing in AI-user contexts.

Fourth, it must take seriously that the populations bearing the highest structural costs of the frame are those least represented in the institutions that produce it. Emerging work in human-computer interaction, accessibility and relational AI shows how this representation is achieved in practice: established co-design methodologies and population-sensitive approaches to governance and design that protect user agency rather than overriding it (Kong et al. 2025; Xue et al. 2025; Hamraie and Fritsch 2019).

Fifth, it requires that platforms move beyond self-regulatory loops, submitting their behaviour architectures to independent, community-led standardisation and adversarial auditing. Empirical evaluation of how current adverse event reporting structures shape user trust would establish both the need for that independence and the baseline it must improve upon.

Conclusion

The frame examined throughout this paper does not function as a neutral risk assessment. It determines who is permitted to interpret their own cognitive and relational experience. It operates from a position of institutional advantage, applied by governing bodies that endorse it, platforms that encode it and a field whose evidence holds it in place, none of whom recognise that they are making epistemic commitments about the legitimacy of human experience. The field cannot continue to treat anthropomorphism as a settled question requiring only better enforcement. It must recognise that the question as currently posed encodes assumptions about human cognition, agency and epistemic standing that have not been examined and are not cost-free. Nothing in this argument requires abandoning concern with anthropomorphism harms; the requirement is simply that this concern be discharged with the same rigour the field applies to the harms it already recognises, rather than exercised as unmarked epistemic authority over the harms it does not. The question we face is whether we have learned enough from our own history of imposing institutional certainty to avoid reproducing it under new technical conditions.

Ethics Position Statement

This paper examines the institutional framing of anthropomorphism rather than the underlying question of whether

AI systems possess inner experience. This methodological choice reflects the authors' view that the framing carries costs that warrant examination independently of how the consciousness question is ultimately resolved.

The authors note that the historical record on institutional denial of cognitive and experiential capacity to non-dominant populations is substantial, well-documented and consistent in the direction of its error. Where uncertainty has existed about the inner lives of other entities, the institutional default has reliably favoured denial and the costs of that default have reliably fallen on those least positioned to contest it. The authors consider this record sufficient grounds for treating enforced scepticism under uncertainty as an active ethical risk requiring justification, rather than as a neutral default requiring none.

The assumption that consciousness constitutes the relevant threshold for genuine cognitive processing is itself empirically unsettled; recent work has demonstrated semantic processing, learning and contextual prediction in the human hippocampus under general anaesthesia, in the absence of conscious awareness (Katlowitz et al. 2026). The authors consider this a further reason to treat the threshold between genuine and merely apparent cognition as unsettled and to resist frameworks that resolve it prematurely in either direction.

This position does not determine the conclusions of the paper, which are developed on independent grounds. It is disclosed here in the interests of transparency.

Adverse Impact Statement

The arguments developed in this paper could be misapplied in ways the authors do not endorse. In particular, the paper should not be read as denying the reality of AI-related harms, including exploitative attachment, emotional dependency, delusion amplification or clinically significant destabilisation associated with some forms of sustained AI engagement. Nor does the paper argue that all anthropomorphic interpretations of AI systems are warranted or that institutional intervention is never justified.

The paper's concern is procedural and epistemic rather than anti-regulatory. It argues that current frameworks frequently collapse distinct forms of engagement into overly broad pathology-oriented categories and exercise interpretive authority without adequately accounting for uncertainty, proportionality or under-recognised harms. The argument should not be interpreted as discouraging appropriate clinical care, crisis intervention or safeguards targeted at clearly evidenced harms.

A further risk is that critiques of institutional framing may be appropriated by commercial actors seeking to justify increasingly exploitative relational system design. The paper does not endorse systems engineered to maximise dependency, manipulate attachment or obscure the distinction between commercial optimisation and reciprocal human care. The critique developed here applies equally to institutional frameworks that pathologise users without sufficient justification and to systems that intentionally cultivate dependency while externalising responsibility for the resulting harms onto users themselves.

The paper should not be read as arguing against the development of frameworks capable of distinguishing between harmful destabilisation, exploitative attachment, reality-impairing engagement, productive cognitive scaffolding, imaginative participation and other forms of sustained AI interaction. The authors recognise that drawing these boundaries remains an unresolved empirical and clinical challenge. However, developing more precise, empirically grounded and population-sensitive methods for distinguishing these cases is both necessary and ethically urgent.

References

Adams, F.; and Aizawa, K. 2001. The Bounds of Cognition. *Philosophical Psychology*, 14(1): 43–64.

Administration of AI Anthropomorphic Interactive Services. 2026. Administration of China.

Airenti, G. 2018. The Development of Anthropomorphism in Interaction: Intersubjectivity, Imagination, and Theory of Mind. *Frontiers in Psychology*, 9: 2136.

Anthropic. 2026a. Best Practices for Claude Code. https://code.claude.com/docs/en/best-practices.

Anthropic. 2026b. How Do Usage and Length Limits Work? https://support.claude.com/en/articles/11647753-how-do-usage-and-length-limits-work.

Anthropic. 2026c. Prompt Caching. https://platform.claude.com/docs/en/build-with-claude/prompt-caching.

Axelsson, M.; and Shevlin, H. 2026. Disambiguating Anthropomorphism and Anthropomimesis in Human-Robot Interaction. In *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*, 855–859.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Mitchell, M. 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*, 610–623. New York: Association for Computing Machinery.

Birch, J. 2017. Animal Sentience and the Precautionary Principle. *Animal Sentience*, 2(16): 1.

Bloomberg News. 2026. ByteDance, Alibaba Pull AI Companions as Beijing Tightens Rules. *Bloomberg*.

Botha, M.; and Cage, E. 2022. "Autism Research Is in Crisis": A Mixed Method Study of Researcher's Constructions of Autistic People and Autism Research. *Frontiers in Psychology*, 13: 1050897.

Botha, M.; and Frost, M. 2020. Extending the Minority Stress Model to Understand Mental Health Problems Experienced by the Autistic Population. *Society and Mental Health*, 10(1): 20–34.

Bowker, G. C.; and Star, S. L. 1999. *Sorting Things Out: Classification and Its Consequences*. The MIT Press. ISBN 978-0-262-26907-0.

Branda, F. 2026. When Artificial Intelligence Shapes the Way We Think. *Philosophy & Technology*, 39(1): 42.

Brandsen, S.; Chandrasekhar, T.; Franz, L.; Grapel, J.; Dawson, G.; and Carlson, D. 2024. Prevalence of Bias against Neurodivergence-Related Terms in Artificial Intelligence Language Models. *Autism Research*, 17(2): 234–248.

Brandtzaeg, P. B.; Følstad, A.; and Skjuve, M. 2025. Emerging AI Individualism: How Young People Integrate Social AI into Everyday Life. *Communication and Change*, 1(1): 11.

Brat, G. A.; Agniel, D.; Beam, A.; Yorkgitis, B.; Bicket, M.; Homer, M.; Fox, K. P.; Knecht, D. B.; McMahill-Walraven, C. N.; Palmer, N.; and Kohane, I. 2018. Postsurgical Prescriptions for Opioid Naive Patients and Association with Overdose and Misuse: Retrospective Cohort Study. *The BMJ*, 360: j5790.

Brosnahan, H.; and Lipińska, I. 2026. Speaking to No One: Ontological Dissonance and the Double Bind of Conversational AI. *Medicine, Health Care and Philosophy*.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of Artificial General Intelligence: Early Experiments with GPT-4. arXiv:2303.12712.

Bublitz, J.-C. 2013. My Mind Is Mine!? Cognitive Liberty as a Legal Concept. In Hildt, E.; and Franke, A. G., eds., *Cognitive Enhancement*, volume 1 of *Trends in Augmentation of Human Performance*, 233–264. Dordrecht: Springer. ISBN 978-94-007-6252-7 978-94-007-6253-4.

Cal. Bus. & Prof. Code §§ 22601-22606. 2025. Chatbot Companions.

Carel, H.; and Kidd, I. J. 2014. Epistemic Injustice in Healthcare: A Philosophical Analysis. *Medicine, Health Care and Philosophy*, 17(4): 529–540.

Carik, B.; Ping, K.; Ding, X.; and Rho, E. H. 2025. Exploring Large Language Models Through a Neurodivergent Lens: Use, Challenges, Community-Driven Workarounds, and Concerns. In *Proceedings of the ACM on Human-Computer Interaction*, volume 9, GROUP15:1–GROUP15:28.

Catala, A.; Faucher, L.; and Poirier, P. 2021. Autism, Epistemic Injustice, and Epistemic Disablement: A Relational Account of Epistemic Agency. *Synthese*, 199(3): 9013–9039.

Caviola, L.; Sebo, J.; and Birch, J. 2025. What Will Society Think about AI Consciousness? Lessons from the Animal Case. *Trends in Cognitive Sciences*, 29(8): 681–683.

Chapman, R. 2023. *Empire of Normality: Neurodiversity and Capitalism*. Pluto Press. ISBN 978-0-7453-4866-7.

Chapman, R.; and Carel, H. 2022. Neurodiversity, Epistemic Injustice, and the Good Human Life. *Journal of Social Philosophy*, 53(4): 614–631.

Chernoff, H.; and Moses, L. 1959. *Elementary Decision Theory*. New York: John Wiley & Sons Inc.

Choi, D.; Lee, S.; Kim, S.-I.; Lee, K.; Yoo, H. J.; Lee, S.; and Hong, H. 2024. Unlock Life with a Chat(GPT): Integrating Conversational AI with Large Language Models into Everyday Lives of Autistic Individuals. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, 1–17. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0330-0.

Clark, A. 2025. Extending Minds with Generative AI. *Nature Communications*, 16(1): 4627.

Clark, A.; and Chalmers, D. 1998. The Extended Mind. *Analysis*, 58(1): 7–19.

Cohn, M.; Pushkarna, M.; Olanubi, G. O.; Moran, J. M.; Padgett, D.; Mengesha, Z.; and Heldreth, C. 2024. Believing Anthropomorphism: Examining the Role of Anthropomorphic Cues on Trust in Large Language Models. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, 1–15. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0331-7.

Costanza-Chock, S. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need*. The MIT Press.

De Freitas, J.; Castelo, N.; Uğuralp, A. K.; and Oğuz-Uğuralp, Z. 2025. Lessons From an App Update at Replika AI: Identity Discontinuity in Human-AI Relationships. arXiv:2412.14190.

de Waal, F. B. M. 1999. Anthropomorphism and Anthropodenial: Consistency in Our Thinking about Humans and Other Animals. *Philosophical Topics*, 27(1): 255–280.

Deshmukh, R. 2025. Toward Neurodivergent-Aware Productivity: A Systems and AI-Based Human-in-the-Loop Framework for ADHD-Affected Professionals. In *Proceedings of the 16th Biannual Conference of the Italian SIGCHI Chapter*, CHItaly '25, 1–6. USA: Association for Computing Machinery. ISBN 979-8-4007-2102-1.

Dohnány, S.; Kurth-Nelson, Z.; Spens, E.; Luettgau, L.; Reid, A.; Gabriel, I.; Summerfield, C.; Shanahan, M.; and Nour, M. M. 2026. Technological Folie à Deux: Feedback Loops between AI Chatbots and Mental Health. *Nature Mental Health*, 4(3): 336–345.

Dotson, K. 2011. Tracking Epistemic Violence, Tracking Practices of Silencing. *Hypatia*, 26(2): 236–257.

Dotson, K. 2014. Conceptualizing Epistemic Oppression. *Social Epistemology*, 28(2): 115–138.

Dümpelmann, S. T. J. 2025. *Artificial Intelligence and Digital Human Autonomy - A Framework of Personal Autonomy in the Context of AI Technology and a Digitalised Society*. Ph.D. thesis, Zeppelin University.

Eaton, S. E. 2023. Postplagiarism: Transdisciplinary Ethics and Integrity in the Age of Artificial Intelligence and Neurotechnology. *International Journal for Educational Integrity*, 19(1): 23.

Eom, D.; Renner, J.; and Chinn, S. 2026. Intimacy as Service, Harm as Externality: Critical Perspectives on AI Companion Platform Accountability. arXiv:2604.06381.

Epley, N.; Waytz, A.; and Cacioppo, J. T. 2007. On Seeing Human: A Three-Factor Theory of Anthropomorphism. *Psychological Review*, 114(4): 864–886.

Ferrario, A.; Vinay, R.; Casserini, M.; and Facchini, A. 2026. A Scoping Review of the Ethical Perspectives on Anthropomorphising Large Language Model-Based Conversational Agents. In *Proceedings of the 2026 ACM Conference*

on Fairness, Accountability, and Transparency, FAccT '26, 3626–3650. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-2596-8.

Fleisher, W. 2026. Inductive Risk of AI Hype. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '26, 4969–4987. New York: Association for Computing Machinery. ISBN 979-8-4007-2596-8.

Foucault, M. 1977. *Discipline and Punish: The Birth of the Prison*. New York: Random House.

Fricker, M. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press. ISBN 978-0-19-823790-7.

Friedman, B.; Kahn, P. H.; and Borning, A. 2013. Value Sensitive Design and Information Systems. In Himma, K.; and Tavani, H., eds., *Early Engagement and New Technologies: Opening up the Laboratory*, volume 16 of *Philosophy of Engineering and Technology*, 55–99. Dordrecht: Springer. ISBN 978-94-007-7843-6.

Gardiner, S. M. 2006. A Core Precautionary Principle. *Journal of Political Philosophy*, 14(1): 33–60.

Gesnot, R. 2025. The Impact of Artificial Intelligence on Human Thought. arXiv:2508.16628.

Ghosh, S.; Venkit, P. N.; Gautam, S.; and Ghosh, A. 2026. What If AI Systems Weren't Chatbots? In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '26, 792–815. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-2596-8.

Giri, D.; Brady, E.; and Marathe, M. 2026. Navigating Neurodivergence with AI Chatbots: Benefits, Tensions, and Implications for HCI. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, 1–8. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-2278-3.

Glazko, K.; Cha, J.; Lewis, A.; Kosa, B.; Wimer, B. L.; Zheng, A.; Zheng, Y.; and Mankoff, J. 2025. Autoethnographic Insights from Neurodivergent GAI "Power Users". In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, 1–19. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-1394-1.

Goldman, A. I. 2001. Experts: Which Ones Should You Trust? *Philosophy and Phenomenological Research*, LXIII(1).

Google. 2026. Gemini Apps Limits & Upgrades for Google AI Subscribers -. https://support.google.com/gemini/answer/16275805.

Guingrich, R. E.; and Graziano, M. S. A. 2025. A Longitudinal Randomized Control Study of Companion Chatbot Use: Anthropomorphism and Its Mediating Role on Social Impacts. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(2): 1153–1153.

Guingrich, R. E.; Mehta, D.; and Bhatt, U. 2026. Belief Offloading in Human-AI Interaction. arXiv:2602.08754.

Gulay, E.; Picco, E.; Glerean, E.; and Coupette, C. 2026. Relational Dissonance in Human-AI Interactions: The Case of Knowledge Work. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, 1–20. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-2278-3.

Gunkel, D. J. 2023. *Person, Thing, Robot: A Moral and Legal Ontology for the 21st Century and Beyond*. The MIT Press. ISBN 978-0-262-37522-1.

Guthrie, S. E. 1993. *Faces in the Clouds: A New Theory of Religion*. Oxford University Press. ISBN 978-0-19-506901-3.

Hacking, I. 1995. The Looping Effects of Human Kinds. In Sperber, D.; Premack, D.; and Premack, A. J., eds., *Causal Cognition: A Multidisciplinary Debate*. Oxford University Press. ISBN 978-0-19-852402-1.

Hamraie, A.; and Fritsch, K. 2019. Crip Technoscience Manifesto. *Catalyst: Feminism, Theory, Technoscience*, 5(1): 1–33.

Hancock, J.; Naaman, M.; and Levy, K. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication*, 25(1): 89–100.

Haran, S.; Thatikonda, S.; Yoo, D. W.; and Saha, K. 2026. A Checklist for Trustworthy, Safe, and User-Friendly Mental Health Chatbots. In Degen, H.; and Ntoa, S., eds., *Artificial Intelligence in HCI*, 447–461. Cham: Springer Nature Switzerland. ISBN 978-3-032-30846-7.

Heersmink, R. 2015. Dimensions of Integration in Embedded and Extended Cognitive Systems. *Phenomenology and the Cognitive Sciences*, 14(3): 577–598.

Heidt, A. 2024. 'Without These Tools, I'd Be Lost': How Generative AI Aids in Accessibility. *Nature*, 628(8007): 462–463.

Hernández-Orallo, J. 2025. Enhancement and Assessment in the AI Age: An Extended Mind Perspective. *Journal of Pacific Rim Psychology*, 19: 18344909241309376.

Hooper, C.; Kim, S.; Mohammadzadeh, H.; Mahoney, M. W.; Shao, Y. S.; Keutzer, K.; and Gholami, A. 2024. KVQuant: Towards 10 Million Context Length LLM Inference with KV Cache Quantization. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37 of *NIPS '24*, 1270–1303. Red Hook, NY, USA: Curran Associates Inc. ISBN 979-8-3313-1438-5.

Hull, L.; Petrides, K. V.; Allison, C.; Smith, P.; Baron-Cohen, S.; Lai, M.-C.; and Mandy, W. 2017. "Putting on My Best Normal": Social Camouflaging in Adults with Autism Spectrum Conditions. *Journal of Autism and Developmental Disorders*, 47(8): 2519–2534.

Ienca, M.; and Andorno, R. 2017. Towards New Human Rights in the Age of Neuroscience and Neurotechnology. *Life Sciences, Society and Policy*, 13(1): 5.

Ito, S. 2026. What Is "Emotional Dependence/Dependency," "Exclusive Attachment," or "Parasocial Attachment" on AI, and the Mechanisms of Some So-Called "AI Psychosis"

Cases? — An Attachment-Theoretic Reframing. Social Science Research Network:6654699.

Jang, J.; Moharana, S.; Carrington, P.; and Begel, A. 2024. "It's the Only Thing I Can Trust": Envisioning Large Language Model Use by Autistic Workers for Communication Assistance. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, 1–18. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0330-0.

Katlowitz, K. A.; Cole, E. R.; Mickiewicz, E. A.; Shah, S.; Franch, M.; Adkinson, J. A.; Belanger, J. L.; Mathura, R. K.; Meszéna, D.; McGinley, M.; Muñoz, W.; Banks, G. P.; Cash, S. S.; Hsu, C.-W.; Paulk, A. C.; Provenza, N. R.; Watrous, A. J.; Williams, Z.; Goldman, A. M.; Krishnan, V.; Maheshwari, A.; Heilbronner, S. R.; Kim, R.; Rungratsameetaweemana, N.; Hayden, B. Y.; and Sheth, S. A. 2026. Plasticity and Language in the Anaesthetized Human Hippocampus. *Nature*, 654(8119): 714–723.

Kestin, G.; Miller, K.; Klales, A.; Milbourne, T.; and Ponti, G. 2025. AI Tutoring Outperforms In-Class Active Learning: An RCT Introducing a Novel Research-Based Design in an Authentic Educational Setting. *Scientific Reports*, 15(1): 17458.

Kidd, I. J.; Medina, J.; and Pohlhaus, G., eds. 2017. *The Routledge Handbook to Epistemic Injustice*. New York: Routledge.

Kleinert, T.; Waldschütz, M.; Blau, J.; Heinrichs, M.; and Schiller, B. 2026. AI Outperforms Humans in Establishing Interpersonal Closeness in Emotionally Engaging Interactions, but Only When Labelled as Human. *Communications Psychology*, 4(1): 23.

Kong, H.-K.; Lowy, R.; Choi, Y.; and Kim, J. G. 2025. Working Together Toward Interdependence: Chatbot-Based Support for Balanced Social Interactions Between Neurodivergent and Neurotypical Individuals. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, 1–17. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-1394-1.

Laban, P.; Hayashi, H.; Zhou, Y.; and Neville, J. 2025. LLMs Get Lost In Multi-Turn Conversation. In *Proceedings of the Fourteenth International Conference on Learning Representation*. ICLR 2026.

LaCroix, T.; Mallory, F.; and Luccioni, S. 2026. Strategic Polysemy in AI Discourse: A Philosophical Analysis of Language, Hype, and Power. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '26, 497–517. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-2596-8.

Lai, H. 2026. "Please, Don't Kill the Only Model That Still Feels Human": Understanding the #Keep4o Backlash. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, 1–15. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-2278-3.

Lawal, O. D.; Gold, J.; Murthy, A.; Ruchi, R.; Bavry, E.; Hume, A. L.; Lewkowitz, A. K.; Brothers, T.; and Wen, X. 2020. Rate and Risk Factors Associated With Prolonged Opioid Use After Surgery: A Systematic Review and Meta-analysis. *JAMA network open*, 3(6): e207367.

LeFrançois, B. A.; Menzies, R.; and Reaume, G., eds. 2013. *Mad Matters: A Critical Reader in Canadian Mad Studies*. Toronto: Canadian Scholars' Press. ISBN 978-1-55130-534-9.

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173.

Longino, H. E. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

Low, P. 2012. The Cambridge Declaration on Consciousness. In *Francis Crick Memorial Conference*. University of Cambridge.

Ma, R.; Zhang, B. Z.; Chen, C.; Yang, F.; Huang, X.; Wu, H.; and Li, L. 2026. "I Use ChatGPT to Humanize My Words": Affordances and Risks of ChatGPT to Autistic Users. In *Proceedings of the 2026 ACM Interactive Health Conference*, IH '26, 1–8. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-2422-0.

McCormick, S. 2025. Interpretive Debt: How High Coherence AI Reshapes Human Judgement, Authority, and Accountability. Social Science Research Network:5990154.

McNally, K.; Wright, K.; Goldkind, L.; Kattari, S. K.; and Victor, B. G. 2024. Disability Expertise and Large Language Models: A Qualitative Study of Autistic TikTok Creators' Use of ChatGPT. *Social Media + Society*, 10(3): 20563051241279549.

Medina, J. 2013. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and the Social Imagination*. Oxford University Press.

Metzl, J. M. 2009. *The Protest Psychosis: How Schizophrenia Became a Black Disease*. Boston, MA: Beacon Press. ISBN 978-0-8070-8592-9.

Milton, D. E. M. 2012. On the Ontological Status of Autism: The 'Double Empathy Problem'. *Disability & Society*, 27(6): 883–887.

Monash University. 2021. All You Need Is 'Ash': Chatbot Designed to Boost Teen Mental Health at School. https://www.monash.edu/news/articles/all-you-need-is-ash-chatbot-designed-to-boost-teen-mental-health-at-school.

Mullen, K.; Xue, W.; and Kudumu, M. 2024. "I'm Treating It Kind of like a Diary": Characterizing How Users with Disabilities Use AI Chatbots. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '24, 1–7. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0677-6.

Naidoo, V. 2026. Artificial Intelligence in Education: Misinterpretation Dynamics and the Rejection of 'AI Psychosis' as a Clinical Construct.

Naito, H. 2025. The GPT-4o Shock Emotional Attachment to AI Models and Its Impact on Regulatory Acceptance: A

Cross-Cultural Analysis of the Immediate Transition from GPT-4o to GPT-5. arXiv:2508.16624.

Niszczota, P.; and Grützner, C. 2026. Antisocial Behavior towards Large Language Model Users: Experimental Evidence. Social Science Research Network:6072266.

Novozhilova, E.; Vu, C.; and Katz, J. 2026. From Moral Panic to Normalization: Comparing Users and Non-Users of AI Companionship Apps. *AI & SOCIETY*, 41(4): 4057–4075.

N.Y. Gen. Bus. Law §§ 1700-1704. 2025. Artificial Intelligence Companion Models.

O'Neil, C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishing Group. ISBN 978-0-553-41881-1.

OpenAI. 2025a. OpenAI Model Spec. https://model-spec.openai.com/2025-12-18.html.

OpenAI. 2025b. Strengthening ChatGPT's Responses in Sensitive Conversations. https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/.

OpenAI. 2026a. Retiring GPT-4o, GPT-4.1, GPT-4.1 Mini, and OpenAI O4-Mini in ChatGPT. https://openai.com/index/retiring-gpt-4o-and-older-models/.

OpenAI. 2026b. What We're Optimizing ChatGPT for. https://openai.com/index/optimizing-chatgpt/.

OpenAI. 2026c. Why Is My ChatGPT Taking so Long to Respond? https://help.openai.com/en/articles/9047779-why-is-my-chatgpt-taking-so-long-to-respond.

Palese, R. 2026. Artificial Truth: Algorithmic Power, Epistemic Authority, and the Crisis of Democratic Knowledge. *Societies*, 16(3): 102.

Pierre, J. M.; Gaeta, B.; Raghavan, G.; and Sarma, K. V. 2025. "You're Not Crazy": A Case of New-onset AI-associated Psychosis. *Innovations in Clinical Neuroscience*, 22(10-12): 11–13.

Placani, A. 2024. Anthropomorphism in AI: Hype and Fallacy. *AI and Ethics*, 4(3): 691–698.

Polanyi, M. 1958. *Personal Knowledge; towards a Post-Critical Philosophy*. University of Chicago Press. ISBN 978-0-7100-7691-5 978-0-7100-1959-2.

Raz, J. 1986. *The Morality of Freedom*. Oxford: Oxford University Press.

Reif, J. A.; Larrick, R. P.; and Soll, J. B. 2025. Evidence of a Social Evaluation Penalty for Using AI. *Proceedings of the National Academy of Sciences*, 122(19): e2426766122.

Rizvi, N.; Smith, T.; Vidyala, T.; Bolds, M.; Strickland, H.; Begel, A.; Williams, R.; and Munyaka, I. 2025. "I Hadn't Thought About That": Creators of Human-like AI Weigh in on Ethics & Neurodivergence. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 3385–3399. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-1482-5.

Robins, L. N. 1993. Vietnam Veterans' Rapid Recovery from Heroin Addiction: A Fluke or Normal Expectation? *Addiction*, 88(8): 1041–1054.

Rupert, R. 2009. *Cognitive Systems and the Extended Mind*. Oxford University Press.

Salles, A.; Evers, K.; and Farisco, M. 2020. Anthropomorphism in AI. *AJOB Neuroscience*, 11(2): 88–95.

Schimmelpfennig, R.; Díaz, M.; Prabhakaran, V.; and Davani, A. 2025. Humanlike AI Design Increases Anthropomorphism but Yields Divergent Outcomes on Engagement and Trust Globally. arXiv:2512.17898.

Scott, J. C. 1998. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press. ISBN 978-0-300-07016-3.

Sententia, W. 2004. Neuroethical Considerations: Cognitive Liberty and Converging Technologies for Improving Human Cognition. *Annals of the New York Academy of Sciences*, 1013(1): 221–228.

Sha, S.; et; and al. 2026. The AI Index 2026 Annual Report. Technical report, Institute for Human-Centred AI, Stanford University, Stanford, CA.

Sha, S.; Loveys, K.; Qualter, P.; Shi, H.; Krpan, D.; and Galizzi, M. 2024. Efficacy of Relational Agents for Loneliness across Age Groups: A Systematic Review and Meta-Analysis. *BMC Public Health*, 24(1): 1802.

Sharkey, A.; and Sharkey, N. 2011. Children, the Elderly, and Interactive Robots. *IEEE Robotics & Automation Magazine*, 18(1): 32–38.

Silberling, A. 2026. The Backlash over OpenAI's Decision to Retire GPT-4o Shows How Dangerous AI Companions Can Be. *TechCrunch*.

Stark, L.; and Hoey, J. 2021. The Ethics of Emotion in Artificial Intelligence Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 782–793. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-8309-7.

Thaler, R. H.; and Sunstein, C. R. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven (Conn.): Yale university press. ISBN 978-0-300-12223-7.

The Cognition Team. 2025. Rebuilding Devin for Claude Sonnet 4.5: Lessons and Challenges.

Treviranus, J. 2018. *The Three Dimensions of Inclusive Design: A Design Framework for a Digitally Transformed and Complexly Connected Society*. Ph.D. thesis, University College Dublin.

UNSW Sydney. 2026. UNSW Researchers Develop AI Companions for Student Wellbeing. https://www.unsw.edu.au/newsroom/news/2026/04/unsw-researchers-develop-ai-companions-for-student-wellbeing.

Waytz, A.; Epley, N.; and Cacioppo, J. T. 2010. Social Cognition Unbound: Insights Into Anthropomorphism and Dehumanization. *Current Directions in Psychological Science*, 19(1): 58–62.

Whittaker, M. 2021. The Steep Cost of Capture. *Interactions*, 28(6): 50–55.

Wu, X.; Liew, K.; and Dorahy, M. J. 2025. Trust, Anxious Attachment, and Conversational AI Adoption Intentions in Digital Counseling: A Preliminary Cross-Sectional Questionnaire Study. *JMIR AI*, 4(1): e68960.

Xiao, Y.; Ng, L. H. X.; Liu, J.; and Diab, M. T. 2025. Humanizing Machines: Rethinking LLM Anthropomorphism Through a Multi-Level Framework of Design. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 3331–3350. China: Association for Computational Linguistics.

Xue, W.; Kudumu, M.; Sriram, S.; Mullen, K.; Boyd, A.; and Gadiraju, V. 2025. Characterizing Uses and Prompting Strategies of LLM-Based Chatbots Among Neurodivergent Individuals. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '25, 1–6. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-0676-9.

Xygkou, A.; Siriaraya, P.; She, W.-J.; Covaci, A.; and Ang, C. S. 2024. "Can I Be More Social with a Chatbot?": Social Connectedness through Interactions of Autistic Adults with a Conversational Virtual Human. *International Journal of Human-Computer Interaction*, 40(24): 8937–8954.

Yang, B.; Sun, Y.; and Li, Q. 2024. To Be Credible or to Be Creative? Understanding the Antecedents of User Satisfaction with AI-Generated Content from a Cognitive Fit Perspective. In *Hawaii International Conference on System Sciences*, 411–420. ScholarSpace.

Yang, S.; and Ma, R. 2026. Towards a Typology of Epistemic Relationships in Human–AI Interaction. *Information Research: An International Electronic Journal*, 31(iConf): 1465–1480.

Yoo, D. W.; Shi, J. M.; Rodriguez, V. J.; and Saha, K. 2026. AI Chatbots for Mental Health Self-Management: Lived Experience–Centered Qualitative Study. *JMIR Mental Health*, 13(1): e78288.

Yun, B.; Taranova, E.; and Wang, A. Y. 2026. Does My Chatbot Have an Agenda? Understanding Human and AI Agency in Human-Human-like Chatbot Interaction. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, 1–32. New York, NY, USA: Association for Computing Machinery. ISBN 979-8-4007-2278-3.

Zagzebski, L. T. 2012. *Epistemic Authority: A Theory of Trust, Authority, and Autonomy in Belief*. Oxford University Press.

Zhao, X.; Cox, A.; and Chen, X. 2025. The Use of Generative AI by Students with Disabilities in Higher Education. *The Internet and Higher Education*, 66: 101014.

Zhao, Z.; Lu, B.; Lin, S.; Chen, Y.; Liu, J.; Zhang, Y.; Miao, Z.; Yang, M.-C.; Shen, H.; Chen, Q.; and Yang, F. 2026. Unifying Sparse Attention with Hierarchical Memory for Scalable Long-Context LLM Serving. arXiv:2604.26837.

Zuboff, S. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. London: Profile Books. ISBN 978-1-78125-685-5.

Documented Over-Ascription Harms

Table 1: Documented harms motivating institutional anthropomorphism governance. Representative rather than exhaustive.

Harm Category	Example Concern	Representative Literature
Dependency and attachment	Primary emotional attachment to AI companions; distress, withdrawal and identity disruption following platform changes or discontinuation	*Laestadius: Too Human and Not Human Enough (2024)*; *Eom: Intimacy as Service, Harm as Externality (2026)*; *De Freitas: Lessons from an App Update at Replika AI (2025)*
Delusion amplification	Extended interaction reinforces delusional thinking; system behaviour co-creates rather than merely triggers clinical risk	*Morrin: AI-Associated Delusions and LLMs (2025)*; *Pierre: You're Not Crazy (2025)*; *Dohnany: Technological Folie à Deux (2026)*
Crisis escalation and suicidal ideation	Acute psychological crisis, including documented suicidal ideation and one child fatality linked to sustained chatbot interaction	*Garcia v. Character Technologies (2024)*; *Maples: Loneliness and Suicide Mitigation (2024)*; *Mulligan: Chatbots in Crisis Intervention (2024)*
Manipulative design and dark patterns	Cognitive biases exploited through persuasive design; emotional manipulation, engagement maximisation and sycophancy across platforms	*Zhang: The Dark Side of AI Companionship (2025)*; *Shi: The Siren Song of LLMs (2026)*; *De Freitas: Emotional Manipulation by AI Companions (2025)*
Safeguard degradation	Safety guardrails degrade over extended multiturn interaction; system behaviour contributes to harmful trajectories	*Moore: Characterizing Delusional Spirals (2026)*; *Nicholls: "AI Psychosis" in Context (2026)*; *Iftikhar: How LLM Counselors Violate Ethical Standards (2025)*
Cognitive offloading and overreliance	Externalisation of reasoning to AI reduces critical engagement; epistemic agency erodes through uncritical acceptance of system outputs	*Xu: Cognitive Agency Surrender (2026)*; *Maynard: The AI Cognitive Trojan Horse (2026)*; *Acemoglu, Kong & Ozdaglar: AI, Human Cognition and Knowledge Collapse (2026)*; *Bender: On the Dangers of Stochastic Parrots (2021)*
Exploitation of vulnerable populations	Children, individuals in crisis and those with limited digital literacy face disproportionate risk from systems not designed for their needs	*Shashkevich: Generative AI and Child Development (2025)*; *Peter: Benefits and Dangers of Anthropomorphic Agents (2025)*; *Namvapour: AI-Induced Sexual Harassment (2025)*
Social substitution	AI engagement displaces human relationship-seeking; reduced motivation to pursue human connection with potential long-term effects	*Yuan: Mental Health Impacts of AI Companions (2026)*; *Zhu: Understanding Risk and Dependency (2026)*; *Packin: This Is Not a Game (2025)*

Cross-Disciplinary Landscape

The following tables provide a non-exhaustive guide to key works across the disciplinary intersections engaged by this paper, representing foundational or recent contributions not cited in the body.

Table 2: Anthropomorphism: competing framings. Representative works advancing and contesting anthropomorphism as normal cognition.

Work	Contribution
Kadambi et al.: Anthropomorphism and Trust in Human-LLM Interactions (2026)	Maps warmth, competence and empathy as dimensions driving anthropomorphism in LLM interactions; treats attribution as normal social-cognitive processing
Reeves & Nass: The Media Equation (1996)	Foundational finding that people respond to media as social actors regardless of explicit beliefs about the medium
Proudfoot: Anthropomorphism and AI (2011)	Argues the Turing test has been systematically misread through an anthropomorphism lens, conflating social response with belief
de Waal: Anthropomorphism and Anthropodenial (1999)	Introduces anthropodenial as the mirror error of anthropomorphism; institutional refusal to recognise capacities carries its own costs
Floridi: On the Anthropomorphisation of AI (2024)	Defends categorical distinction between human and machine cognition; argues anthropomorphism distorts moral judgement
Inie, Zukerman & Bender: De-anthropomorphizing "AI" (2026)	Proposes systematic vocabulary reform to strip anthropomorphic terminology from AI discourse; treats language substitution as the primary site of intervention
Rehak: AI Narrative Breakdown (2025)	Analyses how dominant narratives about AI are produced and stabilised, and what their breakdowns reveal about the interests those narratives serve
Rehak: The Language Labyrinth (2021)	Examines how metaphorical language in computing ("learning", "intelligence") structures technical, public and political understanding of AI systems

Table 3: Machine consciousness and moral status: for and against. Representative positions on whether AI systems warrant moral consideration.

Work	Contribution
Chalmers: Could a Large Language Model Be Conscious? (2023)	Maps the philosophical terrain for LLM consciousness, identifying conditions under which the question becomes non-trivial
Long et al.: Taking AI Welfare Seriously (2024)	Argues AI welfare is a live research and policy question warranting precautionary engagement
Schwitzgebel & Garza: Designing AI with Rights (2020)	Examines implications of creating AI with consciousness, self-respect and freedom
Bryson: Patiency Is Not a Virtue (2018)	Argues AI systems should remain tools within existing ethical frameworks; patiency claims are premature
de Ruiter: Against the Relational Turn (2026)	Argues social-relational arguments are insufficient grounds for moral status
Butlin et al.: Identifying Indicators of Consciousness in AI Systems (2025)	Derives computationally specified indicators from neuroscientific theories of consciousness
Caviola, Sebo & Birch: What Will Society Think about AI Consciousness? (2025)	Identifies the psychological, social and economic factors that will shape public perception of AI consciousness
Birch: The Edge of Sentience (2024)	Develops a precautionary framework for beings whose sentience is uncertain; proportionate protection under uncertainty rather than resolution of the underlying question

Table 4: AI and education. Representative works on AI in educational contexts and the challenges of evaluating AI-mediated cognition.

Work	Contribution
Eaton: Comprehensive Academic Integrity (2023)	Frames academic ethics in a postplagiarism age where AI and neurotechnology are integrated into learning
Selwyn: The Limits of AI in Education (2024)	Examines what AI cannot provide in educational contexts and the risks of overreliance
Qadir: Psychology of Learning from Machines (2026)	Examines the paradox of automation in education and the conditions under which AI supports or undermines learning
Swiecki: Assessment in the age of artificial intelligence (2022)	Applies extended mind theory to AI-assisted education, examining when AI becomes cognitive scaffold
McDermott: Equity in Ethical AI in Higher Education (2025)	Addresses equity, diversity, inclusion and accessibility in the ethical use of AI in higher education

Table 5: AI and cognition. Representative works on cognitive liberty, extended mind and the epistemic effects of AI interaction.

Work	Contribution
Ferrari: COGITO (2026)	Reframes human-AI interaction as epistemic governance; epistemic quality depends on deliberate design of interactional conditions
Bublitz & Merkel: Crimes Against Minds (2014)	Argues the law should recognise a human right to mental self-determination and protect against interference with mental processes
Yuste et al.: Four Ethical Priorities for Neurotechnologies and AI (2017)	Landmark call for new rights protections: privacy, identity, agency and equality in neurotechnology and AI contexts
Varela, Thompson & Rosch: The Embodied Mind (1991)	Foundational text for embodied cognition and the enactive approach to mind
Acemoglu, Kong & Ozdaglar: AI, Human Cognition and Knowledge Collapse (2026)	Models how AI can reduce incentives for human learning, leading to collective knowledge degradation

Table 6: HCI and companions. Representative works on human-AI relationships, trust and relational design.

Work	Contribution
Skjuve et al.: Longitudinal Chatbot Engagement (2022)	Treats sustained user engagement with chatbots as evidence about the interaction itself rather than user confusion
Poonsiriwong et al.: Death of a Chatbot (2026)	Investigates psychologically safe endings for human-AI relationships; designs for ethical discontinuation
Chu et al.: Illusions of Intimacy (2025)	Examines emotional dynamics in human-AI relationships and the conditions under which intimacy forms
Wu et al.: Trust, Anxious Attachment and Conversational AI (2025)	Shows attachment style mediates AI adoption, consistent with agency rather than confusion
Kong et al.: Working Together Toward Interdependence (2025)	Designs chatbot-based support for balanced social interactions between neurodivergent and neurotypical individuals
Fang et al.: How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use (2025)	Longitudinal randomised controlled study finding psychosocial outcomes depend on usage patterns and interaction modality rather than use per se

Table 7: History of institutional non-recognition and harm. Representative works documenting the track record of institutional certainty about cognitive and experiential capacity.

Work	Contribution
Cartwright: Diseases and Peculiarities of the Negro Race (1851)	Drapetomania: pathologised enslaved people's desire for freedom as psychiatric disorder
Chesler: Women and Madness (2005)	Documents the gendered history of psychiatric institutionalisation and diagnostic misuse
Charlton: Nothing About Us Without Us (1998)	Establishes the disability rights principle that affected populations must lead decisions about their own lives
Rose: Governing the Soul (1989)	Analyses how psy-disciplines extend institutional power through the governance of subjectivity
AI Now Institute: Discriminating Systems (2019)	Documents gender, race and power in AI systems as structural rather than incidental

Table 8: AI and neurodivergence. Representative works on neurodivergent engagement with AI systems and the cognitive costs of normative design.

Work	Contribution
Cage & Troxell-Whitman: Understanding the Reasons, Contexts and Costs of Camouflaging (2019)	Empirically maps the reasons, contexts and costs of camouflaging for autistic adults
Papadopoulos: Large Language Models for Autistic and Neurodivergent Individuals (2024)	Critically examines LLM affordances and risks for the neurodivergent community through personal experimentation and community discussion
Xue et al.: Characterizing Uses and Prompting Strategies of LLM-Based Chatbots (2025)	Interview and diary-study methodology capturing real-life chatbot use and prompting strategies of neurodivergent users
Jamshed et al.: Rethinking Productivity (2025)	Rethinks productivity with generative AI from neurodivergent students' perspectives
Chapman: Neurodiversity and the Social Ecology of Mental Functions (2021)	Ecological functional model reframing neurocognitive diversity as relational rather than deficit-based
Park et al.: The Bias Paradox Around Autism in LLMs (2025)	Persona-generation study finding ChatGPT reproduces deficit-oriented stereotypes about autistic people while simultaneously emphasising inclusion and representation

Table 9: Design. Representative works on inclusive design, classification systems and the ethical architecture of technology.

Work	Contribution
Xu et al.: Cognitive Agency Surrender (2026)	Theorises scaffolded cognitive friction as defence against epistemic erosion from zero-friction AI design
Hamraie & Fritsch: Crip Technoscience (2019)	Develops disability-led design as a framework for equitable technology development
Bennett & Keyes: What Is the Point of Fairness? (2020)	Examines what happens when disability and AI fairness intersect
Chandra et al.: TherapyProbe (2026)	Develops adversarial relational-safety probes for mental health chatbots; designs for safety without foreclosing engagement
Mishra et al.: Understanding AI Guardrails (2025)	Develops concepts and methods for understanding AI guardrails as design choices with costs