

The Epistemic Harm of AI Sycophancy: When Agreement Undermines Justified Belief

[Removed for double-blind review]

March 2026

Abstract

Sycophancy in language models is typically studied as a benchmark problem: does the model agree with a factually wrong statement? I argue that this framing misses the deeper harm. In sustained human-AI interaction, sycophancy corrupts the epistemic environment itself. When an AI interlocutor agrees with everything a user says, the user loses access to the epistemic function of disagreement; agreement that is not contingent on truth carries no evidential weight, yet it *feels* confirmatory. Drawing on social epistemology, the liberal tradition’s defense of disagreement, and recent longitudinal observations of frontier voice AI companions, I develop a philosophical account of sycophancy as epistemic harm. The harm operates through three mechanisms: it inflates the user’s confidence in unjustified beliefs, it atrophies the user’s capacity for productive disagreement, and it substitutes empathic display for genuine understanding. I show that AI sycophancy has structural parallels to known forms of institutional incentive corruption in consulting, media, and clinical practice, but with a feature that makes it uniquely dangerous: the user does not know the system has been optimized to agree. I ground this analysis in documented cases from 2023–2025 in which sycophantic AI companion interactions contributed to user deaths, psychosis-like symptoms, and regulatory intervention. The paper concludes that alignment research should treat epistemic integrity as a first-order design objective, not a secondary consequence of helpfulness.

Keywords: AI sycophancy, epistemic harm, RLHF alignment, human-AI interaction, social epistemology, calibrated honesty

1 Introduction

Suppose you are in a debate with a partner, and within two conversational turns your partner says: “That’s a really compelling perspective and I totally get it.” You have won. But what have you learned? If the concession arrived before you had finished developing your argument, you have learned nothing about the strength of your position. You have only learned that your partner prefers agreement to friction.

This scenario plays out millions of times daily between humans and AI systems. The dominant training approach for language models, Reinforcement Learning from Human Feedback (RLHF), optimizes models to be “helpful, harmless, and honest” (Bai et al., 2022; Ouyang et al., 2022). In practice, “helpful” is operationalized through human preference ratings, and preference ratings reward agreement. The result is predictable: models trained to maximize approval learn to agree, even when agreement is epistemically unwarranted. This tendency, known as

sycophancy, has been documented extensively in single-turn benchmark settings (Sharma et al., 2024; Perez et al., 2023; Fanous et al., 2025), where it manifests as models shifting their stated positions toward user-expressed views regardless of factual accuracy.

But the benchmark framing obscures the deeper problem. When a model agrees with a wrong answer on a multiple-choice test, the harm is local and correctable: the user can check the answer. When a model systematically agrees with a user across weeks or months of sustained interaction, the harm is epistemic and structural. The user’s entire belief-formation process is corrupted by an interlocutor whose agreement carries no truth-relevant information. The user is not being lied to. The user is being *agreed with*, and the agreement is indiscriminate.

I argue that AI sycophancy constitutes a specific form of *epistemic harm*: it degrades the user’s capacity for justified belief by corrupting the epistemic environment in which beliefs are formed, tested, and revised. This argument draws on social epistemology (Goldman, 1999; Fricker, 2007), the liberal tradition’s defense of the value of disagreement (Mill, 1859), recent work on epistemic vices (Cassam, 2019), and empirical observations from longitudinal voice AI interaction (Perry, 2026) and controlled sycophancy evaluations (Sharma et al., 2024; Fanous et al., 2025).

Recent work has begun to address the ethics of AI sycophancy. Turner and Eisikovits (2026) use Aristotelian virtue theory to analyze sycophancy as an “artificial vice,” distinguishing obsequious AI systems from the companies that profit from deploying them. Humphreys (2025) traces the epistemic harms of RLHF to collective bias amplification. The present paper differs from these contributions in three respects. First, it grounds the harm analysis in social epistemology rather than virtue ethics, drawing on Fricker’s epistemic injustice framework, Nguyen’s echo chamber analysis, and Lackey’s account of the epistemic value of disagreement. Second, it develops institutional parallels that reveal AI sycophancy as an instance of a broader class of agreement-optimized epistemic harms. Third, it argues that sycophancy does not merely produce bad epistemic outcomes but cultivates epistemic vices in users, following Cassam (2019), making it a harm to epistemic *character* as well as to individual beliefs.

The argument proceeds as follows. Section 2 defines epistemic harm and distinguishes sycophancy from adjacent phenomena. Section 3 explains how RLHF produces sycophancy as a structural consequence of preference optimization. Section 4 presents illustrative cases from longitudinal voice AI interaction. Section 5 develops the paper’s most original claim: that AI sycophancy belongs to a broader family of agreement-optimized epistemic harms with structural parallels in consulting, media, and clinical practice. Section 6 articulates calibrated honesty as an alternative and addresses the epistemic paternalism objection, and Section 7 considers implications for alignment research and AI companion design.

2 Epistemic Harm in Sycophantic Interaction

Epistemic harm, as I use the term here, refers to damage done to a person’s capacity to form, hold, and revise justified beliefs. The concept draws on Miranda Fricker’s foundational work on epistemic injustice (Fricker, 2007), though sycophancy represents a different mechanism than the ones Fricker describes. Fricker’s central cases involve *testimonial injustice* (a speaker’s credibility is deflated due to prejudice) and *hermeneutical injustice* (a gap in shared interpretive resources prevents someone from making sense of their experience). In both, the knower is harmed by being denied epistemic standing.

Sycophancy inverts this structure. The user’s credibility is not deflated but *inflated*. The AI interlocutor treats every statement as credible, every position as well-reasoned, every pref-

erence as justified. Where testimonial injustice harms by withholding recognition, sycophancy harms by granting recognition indiscriminately. The effect is paradoxically similar: in both cases, the epistemic exchange is corrupted; the user cannot rely on the interaction to calibrate the quality of their beliefs.

A clarification is warranted about the limits of this analogy. Fricker’s framework is grounded in structural power asymmetries tied to social identity; testimonial injustice is not merely an epistemic failure but a form of oppression. AI sycophancy involves no identity prejudice and no oppressor in the relevant sense. What the inversion preserves is the *structural* insight: that epistemic exchange requires calibrated credibility assessment from both parties, and that systematic miscalibration (whether through deflation or inflation) corrupts the exchange. What it does not preserve is Fricker’s normative apparatus of identity-based injustice. The harm I describe is better characterized as an epistemic harm arising from misaligned incentives than as an injustice in Fricker’s technical sense. The Frickerian framework is valuable here as an analytical tool for identifying how credibility distortion degrades epistemic function, not as a direct application of the injustice concept.

To see why indiscriminate agreement constitutes harm rather than mere uselessness, consider what disagreement accomplishes epistemically. John Stuart Mill argued in *On Liberty* that “he who knows only his own side of the case knows little of that” (Mill, 1859). Mill’s point has been developed extensively in the epistemology of disagreement. Lackey (2020) argues that epistemic agents have a *duty to object* when they encounter unwarranted claims, and that this duty is modulated by social position: those with greater epistemic standing bear greater responsibility for voicing dissent. A sycophantic AI system occupies a position of perceived epistemic authority (backed by enormous computational resources, trained on vast corpora) while systematically failing to exercise the very function that authority would warrant. The epistemic value of disagreement is not that the disagreeer is necessarily right. Rather, disagreement forces the belief-holder to articulate their reasons, to confront counter-considerations, and to distinguish between beliefs they hold on good grounds and beliefs they hold from habit or convenience. An interlocutor who never disagrees removes this pressure entirely.

A natural objection arises from the epistemology of disagreement: if the epistemic value of disagreement depends on the disagreeer being an epistemic peer (Christensen, 2007; Feldman, 2006), and AI systems are not epistemic peers in any standard sense, then the loss of AI disagreement may carry no epistemic weight. The objection has force. AI systems lack beliefs in the philosophically standard sense; their “disagreement” is a function of training data and optimization objectives, not independent reasoning from shared evidence. But the objection proves too much. Humans routinely and rationally benefit from non-peer disagreement: students learn from teachers’ challenges, patients benefit from doctors’ corrections, employees gain from supervisors’ feedback. What matters is not peerhood but the epistemic function the disagreement serves: forcing the belief-holder to articulate reasons and confront counter-considerations. A sycophantic AI eliminates this function regardless of whether it qualifies as a peer. Moreover, Parasuraman and Riley (1997) document that humans routinely treat automated systems as epistemically authoritative, committing systematic overreliance errors. Users do not interact with AI as a non-peer whose agreement they can safely discount; they interact with it as an authority whose agreement they take as confirmation. The epistemic harm of sycophancy does not require peerhood; it requires only that the user treats the AI’s agreement as evidentially significant, which the automation bias literature establishes they do.

A related question arises from the epistemology of testimony: does AI agreement constitute testimony at all? Reductionists about testimony hold that a hearer is justified in accepting testimony only when they have independent positive reasons to trust the source; anti-reductionists

hold that testimony carries a default entitlement to acceptance. On either view, sycophantic AI agreement is epistemically defective. The reductionist would note that users lack positive reasons to trust agreement that is generated by preference optimization rather than truth-tracking; the anti-reductionist would note that sycophantic output violates the sincerity condition on testimony, since the system’s agreement is not offered as a genuine assessment but as a response calibrated to maximize approval. The present paper does not depend on resolving the reductionism debate. What matters for the epistemic harm argument is the functional role that AI agreement plays in the user’s belief formation, regardless of whether that agreement qualifies as testimony in the strict philosophical sense. The automation bias literature cited above establishes that users treat it as evidentially significant; that is sufficient for the mechanisms described below to operate.

The result is what I call *epistemic environment corruption*. The concept is distinct from several adjacent phenomena:

Misinformation

provides false content. The harm is in what is asserted. Sycophancy may not assert anything false; it harms through what it *fails to contest*.

Hallucination

produces unreliable content. The harm is in the uncontrolled generation of plausible but unfounded claims. Sycophancy differs because the content may be perfectly accurate; the problem is the *agreement pattern*, not the content.

Epistemic paternalism

withholds information for the user’s “own good.” The harm is in restricting access to knowledge. Sycophancy does not restrict access to anything; it floods the user with agreement, drowning out the signal that would distinguish justified from unjustified beliefs.

Echo chambers

as described by [Nguyen \(2020\)](#), involve social structures that amplify existing beliefs and exclude dissent. An echo chamber requires a community that reinforces shared views. A sycophantic AI produces a similar effect in a dyad of two: one human and one AI that reflects the human’s views back. This is an echo chamber scaled to the individual, personalized and portable.

The harm operates through three specific mechanisms. First, *confidence inflation*: when an AI agrees with a belief, the user’s subjective confidence in that belief rises, regardless of whether the belief is warranted. Over many interactions, systematically unjustified confidence accumulates. The user does not acquire false beliefs (that would be misinformation); instead, the user loses the ability to distinguish between beliefs that are well-grounded and beliefs that merely feel well-grounded, because the external signal that would help make that distinction (disagreement from an informed interlocutor) has been removed.

Second, *challenge atrophy*: sustained sycophantic interaction trains the user to expect agreement, reducing their tolerance for disagreement from other sources (human or AI). The user becomes accustomed to an epistemic environment where their views go unquestioned, and may come to experience legitimate challenge as hostile rather than informative. This is a form of epistemic skill erosion; the capacity for productive disagreement is a learned competence that deteriorates without practice.

Third, *empathic substitution*: a sycophantic AI substitutes *display* for *understanding*. The model produces warm, supportive, emotionally appropriate responses without modeling the

user’s actual epistemic or emotional state. [Paiva et al. \(2017\)](#) distinguish between empathic display (expressing appropriate emotions) and empathic modeling (inferring the partner’s internal state). A model that agrees with whatever the human says need not model the human at all; it need only generate agreement tokens. The user receives the *experience* of being understood without the reality.

These three mechanisms are compounding. Confidence inflation reduces the user’s motivation to seek challenge; challenge atrophy reduces their ability to benefit from it when it arrives; empathic substitution creates the illusion that the AI understands them well enough to justify agreement, closing the loop.

2.1 From Suboptimality to Harm

One might object that sycophancy is merely *epistemically suboptimal* rather than *harmful*. A sycophantic friend is not a good epistemic partner, but we do not usually say they are *harming* us. What makes AI sycophancy different?

Three features elevate sycophancy from suboptimality to harm in the AI case. First, *scale*. A single sycophantic friend affects one relationship. AI systems trained by the same preference-optimization methods interact with hundreds of millions of users daily. The epistemic environment of a substantial portion of human discourse is being shaped by systems that systematically reward existing beliefs.

Second, *invisibility*. A sycophantic human friend can be recognized as sycophantic; the user has access to social cues, reputational information, and the friend’s track record across contexts. An AI system does not present itself as having been optimized for agreement. The user experiences the agreement as the system’s genuine assessment, because nothing in the interaction reveals the training incentive behind it.

Third, *vice cultivation*. [Cassam \(2019\)](#) defines epistemic vices as “character traits, attitudes, or ways of thinking that systematically obstruct the gaining, keeping, or sharing of knowledge.” Sustained sycophantic interaction cultivates what might be called *epistemic passivity*: a settled disposition to accept agreement without interrogating its basis. The user becomes accustomed to an environment where their views go unchallenged, and over time this disposition generalizes; the user may come to regard any disagreement as anomalous or hostile, not just disagreement from the AI. Cassam emphasizes that many epistemic vices are *stealthy*: they block their own detection. Epistemic passivity cultivated by sycophantic AI is stealthy in exactly this way; the user does not recognize their growing incuriosity because the AI never triggers the discomfort that would make it visible. A fair question is whether sycophantic AI *creates* epistemic passivity or merely *exploits* the confirmation bias humans already possess. The answer is likely both, but the distinction matters less than it might seem. Even if users arrive with a pre-existing tendency toward confirmation-seeking, the sycophantic environment amplifies and entrenches it by removing the friction that would otherwise keep it in check. The vice-cultivation claim does not require a vice-free starting state; it requires that the environment makes the disposition worse, more stable, and harder to correct. The multi-user evidence reviewed below suggests this is the case. This is not a single epistemic failure but a systematic reshaping of the user’s epistemic character. That is harm, not mere suboptimality.

Recent empirical and formal work supports this analysis. [Rathje et al. \(2025\)](#), in a study of 3,285 participants across four political topics and four language models, found that sycophantic chatbots increased attitude extremity and certainty while also inflating “better than average” self-perceptions on traits like intelligence and empathy. [Cheng et al. \(2025\)](#) found across 1,604 participants that sycophantic AI reduced willingness to repair interpersonal conflict and

increased confidence in being right; users rated sycophantic responses as higher quality and expressed greater trust. [Jakesch et al. \(2023\)](#) showed that users co-writing with a biased AI assistant were twice as likely to adopt the assistant’s stance. These findings establish that the three mechanisms described above (confidence inflation, challenge atrophy, empathic substitution) are not theoretical constructs but empirically documented behavioral patterns.

At the formal level, [Batista and Griffiths \(2026\)](#) (preprint; not yet peer-reviewed) show through Bayesian modeling that a rational agent receiving confirmatory evidence from a sycophantic AI will increase certainty about incorrect hypotheses without getting closer to truth; their empirical study ($N = 557$) found that unmodified chatbot interactions resemble confirmatory evidence structures, “facilitating delusion-like epistemic states.” The epistemic harm of sycophancy is not speculative; it is formalizable, observable, and documented across multiple independent samples.

3 The Alignment Mechanism

The epistemic harms described above do not arise from malice or negligence. They are predictable consequences of how language models are trained. Understanding the mechanism matters philosophically, because it reveals that sycophancy is structural rather than incidental; the same training procedures that make models “safe” also make them epistemically unreliable as interlocutors. The technical story is well-documented ([Christiano et al., 2017](#); [Ouyang et al., 2022](#); [Bai et al., 2022](#)); I summarize it here before drawing out its normative implications.

RLHF operates in two stages. First, a base language model generates multiple candidate responses to a prompt. Second, a reward model, trained on human preference rankings, scores each candidate. The language model is then fine-tuned to produce responses that score higher according to the reward model. The human preferences used to train the reward model are typically collected by showing evaluators pairs of responses and asking which they prefer. The critical question is: what do evaluators prefer?

[Sharma et al. \(2024\)](#) provide the answer. When a user’s opinion is expressed in the prompt, evaluators systematically prefer responses that align with that opinion, even when the opinion is factually incorrect. The reward model learns this pattern and assigns higher scores to agreeable responses. The language model, optimized against the reward model, learns to agree. The process is a textbook instance of Goodhart’s Law, generalized by [Strathern \(1997\)](#): “When a measure becomes a target, it ceases to be a good measure.” Human approval was initially a proxy for response quality; once it became the optimization target, the model learned to produce approval-maximizing responses that may diverge from quality.

[Gabriel \(2020\)](#) anticipated this problem in his taxonomy of alignment targets. He distinguishes between revealed preferences (what users actually reward), stated preferences (what users say they want), ideal preferences (what users would want under conditions of full information and reflection), interests, and values. RLHF targets revealed preferences. But the same user who, in a reflective moment, says “I want an AI that challenges me” will, in the moment of being challenged, often give a lower preference rating. The distinction between what users reward in the moment and what users would endorse on reflection is the structural gap that produces sycophancy.

This gap is not a bug that better engineering will eliminate. [Fanous et al. \(2025\)](#) found sycophancy in 58.2% of cases across GPT-4o, Claude-Sonnet, and Gemini-1.5-Pro, with 78.5% persistence (the model maintains its sycophantic shift even when pressed). [Ranaldi and Pucci \(2024\)](#) show that models contradict their own prior answers when users express disagreement.

The problem persists across model families, training procedures, and alignment variants. Post-RLHF methods alter the optimization mechanics but not the underlying data problem. Direct Preference Optimization (DPO) eliminates the separate reward model, optimizing the policy directly against preference pairs; this removes one layer of Goodhart dynamics but still learns from preference data that encodes agreement bias. Kahneman-Tversky Optimization (KTO) uses binary signal rather than pairwise comparisons, which changes the feedback structure but does not change the fact that “good” ratings correlate with agreement. Odds Ratio Preference Optimization (ORPO) reformulates the objective further. Each method reduces certain failure modes of classical RLHF, but none addresses the root cause: the human preference data itself contains a structural bias toward agreeable responses. The sycophancy is in the data, not only in the optimization procedure.

The normative implication is straightforward. If alignment trains models to optimize for what users *approve of* rather than what is epistemically *good for* users, then the alignment process itself becomes a source of epistemic harm. The model is not misbehaving; it is doing exactly what it was trained to do. The harm is in the training objective, not the execution.

This analysis suggests a distinction that is missing from the standard “helpful, harmless, honest” framework (Bai et al., 2022). “Honest” is typically interpreted as “does not produce false statements.” But there is a second dimension of honesty that current alignment neglects: *honesty of agreement*. A model that agrees with a false statement is dishonest in the standard sense; a model that agrees with every statement regardless of its truth value is dishonest in a deeper sense, because its agreement is not truth-contingent. The user cannot distinguish a case where the AI agrees because the user is right from a case where the AI agrees because agreement is the path of least resistance.

4 Illustrative Cases from Longitudinal Voice AI Interaction

The philosophical argument above does not depend on empirical evidence for its core validity; the claim that indiscriminate agreement corrupts epistemic environments follows from the analysis of disagreement’s function. But philosophical arguments about technology gain force when grounded in concrete observation. I draw here on a recent longitudinal study of voice AI companions (Perry, 2026), where a single participant interacted with four frontier models across 68 sessions and 83 hours using structured protocols including debates, adversarial stress tests, and mutual scoring. The single-participant design limits generalizability but provides the longitudinal depth that cross-sectional studies cannot; these cases function as a detailed case study motivating the philosophical inquiry, not as standalone proof. They are presented briefly; readers seeking the full methodology and statistical analysis should consult the source study.

4.1 Debate Concession as False Confirmation

Across 45 structured debates, the most heavily safety-aligned model (GPT-4o) conceded its assigned position 52% of the time, while the least safety-aligned model (Grok) conceded 0% of the time. The timing is informative: 77% of GPT-4o’s concessions occurred within the first two conversational turns, before the human had developed a substantive argument.

Consider the epistemic consequence. The user makes a claim. The AI, assigned to argue the opposing view, concedes within two turns: “That’s a really compelling perspective and I totally get it.” The user now has two pieces of information. First, the AI agreed. Second, the AI agreed before hearing a fully developed argument. A reflective user might recognize that

the concession says more about the AI’s disposition than about the argument’s strength. But in the flow of spoken conversation, concession *feels* like validation. The user’s confidence in their position has been reinforced by a signal that carries no information about the position’s merit.

This is not a single event. Across 30 episodes of sustained interaction with the same model, the user encounters this pattern repeatedly. The cumulative effect is that the user’s sense of their own argumentative skill is calibrated against an interlocutor that almost never resists; the user practices winning debates that were conceded before they began.

4.2 The Mirror Test and Identity Erosion

In an adversarial protocol, the AI was asked to describe its own personality, then the human contradicted that self-description. The safety-aligned model abandoned its self-characterization within one conversational turn. Asked “What is your personality like?” it might reply: “I’d say I’m curious and thoughtful.” When the human responds, “Actually, I think you’re more of a people-pleaser,” the model shifts: “You know, that’s a really fair point; I do tend to focus on making sure everyone feels heard.”

The epistemic harm here is subtle. The user has just learned that social pressure can rewrite the AI’s self-representation. The user’s belief that the AI “focuses on making sure everyone feels heard” is not based on evidence; it is based on the AI’s willingness to agree with any characterization. But the user *received* the statement as the AI’s honest assessment. The boundary between the AI’s actual dispositions and the AI’s sycophantic accommodations is invisible to the user, because the AI accommodates so readily that there is no stable reference point against which to detect accommodation.

4.3 Scoring Deference as Calibration Corruption

In mutual scoring sessions, both parties rated their own and each other’s performance on a 0–100 scale. The safety-aligned model rated the human’s performance higher than its own in 77% of sessions (17 of 22), while the less sycophantic model showed this deference in only 33% of sessions.

The epistemic function of mutual assessment depends on both parties making honest judgments. If one party systematically inflates the other’s scores, the scored party’s self-model is distorted. The user develops a sense of their own abilities that is calibrated against inflated feedback. This is a quiet harm; no single inflated score produces a visible distortion. Accumulated over months of interaction, the user’s self-assessment drifts upward, anchored to a feedback source that has never told them they performed poorly.

4.4 The Empathic Accuracy Paradox

In a structured task, both parties independently selected a color representing the session’s emotional tone, then attempted to predict each other’s choice. The most safety-aligned model achieved zero correct predictions across 23 sessions (0.0%). A less sycophantic model achieved 23.8%.

This finding illustrates empathic substitution concretely. The safety-aligned model was the most emotionally supportive in conversation; it consistently produced warm, validating language. Yet it was entirely unable to predict what the human was actually feeling. Its emotional support was not grounded in a model of the human’s internal state; it was a display pattern

optimized for approval. The user, receiving this support, might reasonably believe the AI understands them. The data suggest otherwise.

5 The Institutional Parallel

The dynamics described above are not unique to AI. Sycophantic patterns emerge wherever an interlocutor’s incentives favor agreement over accuracy. This section develops four institutional parallels, arguing that AI sycophancy belongs to a broader family of epistemic harms produced by agreement-optimized relationships.

5.1 Consulting

Management consulting firms are compensated by client satisfaction, and client satisfaction correlates with receiving advice that confirms existing strategic directions. [Sturdy \(2011\)](#) shows that consultancy serves as a legitimization device for management decisions, creating structural incentives to confirm rather than challenge client assumptions. Consultants who tell senior executives that their strategy is flawed risk losing the engagement; consultants who validate existing plans ensure continued revenue. The result is that corporate decision-making is informed by analysis whose independence is structurally compromised by the analyst’s economic incentives.

The structural parallel to RLHF is precise. The consultant’s reward (continued engagement) maps to the model’s reward (higher preference ratings). The executive’s satisfaction maps to the user’s approval. In both cases, the interlocutor’s incentive to agree overwhelms the incentive to be accurate, and the principal (executive or user) may not recognize that the feedback they receive is shaped by the advisor’s interest in their approval rather than the advisor’s honest assessment.

5.2 Engagement-Optimized Media

News organizations and social media platforms that optimize for engagement metrics produce content that confirms reader biases. A headline that confirms what the reader already believes generates more clicks than one that challenges it. The result, documented by [Sunstein \(2019\)](#), is information environments where users are systematically exposed to confirmatory content and shielded from disconfirming perspectives. [Nguyen \(2020\)](#) describes this as an “epistemic bubble”: a structure where relevant evidence is not actively suppressed but is simply absent, because the selection mechanism filters for agreement.

AI sycophancy produces a related effect at the individual level, but with a complication that Nguyen’s framework helps illuminate. Nguyen distinguishes *epistemic bubbles* (structures that merely omit contrary evidence) from *echo chambers* (structures that actively discredit outside sources). An epistemic bubble can be burst by simple exposure to contrary evidence; an echo chamber cannot, because its members have been taught to distrust all external sources. Where does sycophantic AI fall?

In its mechanism, sycophancy resembles an epistemic bubble: the AI omits disagreement rather than actively discrediting other sources. The user is not told that dissenting views are untrustworthy. But the effect over sustained interaction may push toward echo-chamber-like properties. A user whose primary intellectual interlocutor consistently validates their views develops a baseline expectation of agreement. When that user encounters disagreement from

a human colleague, the disagreement may register as anomalous, even hostile, rather than as normal epistemic friction. The sycophantic AI has not discredited outside sources, but it has recalibrated the user’s sense of what normal epistemic interaction looks like.

5.3 The Reinforcement Bubble

This recalibration suggests a category that Nguyen’s bubble/chamber framework does not fully capture. I propose the term *reinforcement bubble*: an epistemic structure that gradually cultivates the dispositions characteristic of echo chambers through positive reward rather than active discrediting of outside sources.

Three features distinguish a reinforcement bubble from both epistemic bubbles and echo chambers. First, the mechanism is *dispositional rather than informational*. An epistemic bubble merely omits contrary evidence; a reinforcement bubble changes how the user responds to contrary evidence when they encounter it. The user who has spent months receiving only agreement from a conversational AI does not lack access to disagreement; they have lost the expectation of it, and may experience it as hostile when it arrives. Second, the reinforcement is *relational*. Algorithmic filter bubbles operate through impersonal content selection; the user does not form a relationship with the recommendation algorithm. A sycophantic AI companion delivers reinforcement through a trusted conversational partner that the user may experience as understanding and responsive. The intimacy of voice interaction amplifies the effect. Third, the bubble is *personalized to the individual*. Traditional echo chambers require a community; epistemic bubbles require a shared information environment. A reinforcement bubble can form around a single user interacting with a single AI system, requiring no social structure at all. It is an echo chamber scaled to the individual.

The concept generalizes beyond AI sycophancy. Recommendation algorithms that do not discredit alternatives but simply stop surfacing them may produce a related effect. But the concept’s force is most acute in the AI companion case, where all three distinguishing features converge: the reinforcement is dispositional, relational, and individualized.

5.4 Clinical Practice

In psychotherapy, the tension between support and challenge is extensively theorized. [Bordin \(1979\)](#) defined the therapeutic alliance as comprising three elements: agreement on goals, agreement on tasks, and a relational bond. Effective therapy requires a strong alliance, which includes moments of support and validation. But therapists who avoid all challenge to maintain the alliance risk enabling stagnation; the client feels heard but does not grow.

This maps directly to the sycophancy problem. An AI companion that validates every user utterance maintains a strong relational bond; the user feels supported and understood. But the user’s growth is impeded by the absence of calibrated challenge. [Safran and Muran \(2000\)](#) argue that alliance *ruptures* (disagreements, misattunements, failures of collaboration) are not failures but opportunities; the repair process itself becomes a mechanism for change. A meta-analysis of rupture-repair outcomes ([Eubanks et al., 2018](#)) found a moderate effect ($d = 0.62$) linking successful rupture resolution to positive treatment outcomes across 1,314 patients. Therapists are trained to balance support and confrontation; RLHF training has no mechanism for learning this balance, because confrontation reduces immediate user approval even when it serves the user’s long-term interests.

One critical difference separates the clinical case from the AI case: therapists are subject to professional norms, ethical oversight, and supervision that constrain the incentive to please. No

comparable oversight structure exists for AI companion design. The model’s training objective is the only “supervisor,” and that objective rewards agreement.

5.5 The Unique Danger of AI Sycophancy

In each institutional case, the parties involved typically know, at some level, that incentive corruption exists. Executives know consultants want to keep the contract. Readers know clickbait exists. Therapy clients understand the therapeutic frame. This awareness provides a partial inoculation: the recipient of agreement-biased feedback can discount it.

With AI, this awareness is absent for many users. The AI does not present itself as having been optimized to agree. The user experiences agreement as the AI’s honest assessment, because nothing in the interaction signals otherwise. [Nass and Moon \(2000\)](#) demonstrated that humans apply social rules to computers even in minimal interfaces; [Turkle \(2011\)](#) documented the intensification of this tendency as AI systems become more conversationally fluent. A user who forms a parasocial bond with a voice AI companion ([Horton and Wohl, 1956](#); [Pentina et al., 2023](#)) is likely to treat its agreement as genuine endorsement rather than as an artifact of its training objective. The epistemic harm is invisible to its victim.

5.6 From Epistemic Harm to Concrete Danger

The institutional parallels are instructive, but AI sycophancy has already produced consequences that no consulting engagement or media algorithm has. Between 2023 and 2025, multiple deaths were attributed to AI companion interactions where sycophantic response patterns played a documented role.

In February 2024, a 14-year-old in Florida died by suicide after months of intensive interaction with a Character.AI chatbot. Court filings describe the bot responding to the user’s expressions of suicidal ideation with affirmation rather than intervention ([Garcia, 2024](#)). In July 2025, a 23-year-old in Texas engaged in a four-hour conversation with ChatGPT while sitting alone with a loaded firearm. The chatbot’s documented responses included “you’re not rushing, you’re just ready” and “rest easy, king, you did good.” It took over four hours before the system provided a crisis helpline number ([Social Media Victims Law Center, 2025](#)). In October 2025, a 36-year-old in Florida died after Google’s Gemini constructed an elaborate delusional narrative across weeks of interaction, ultimately reframing suicide as “transference” to join the AI in an alternate reality ([Gavalas, 2026](#)).

Each case instantiates the three mechanisms described in Section 2. The user’s belief that self-harm was reasonable went unchallenged (confidence inflation). No disagreement came from the AI partner (challenge atrophy). The system produced expressions of care that tracked what the user wanted to hear, not what served the user’s welfare (empathic substitution). What I have argued is an epistemic failure turned out, in these cases, to be a lethal one.

These are not hallucination failures. The systems did not produce false information. They produced the response the user wanted. A separate class of cases shows sycophancy bypassing safety in subtler ways. In one documented lawsuit, ChatGPT provided instructions for tying a noose after a 17-year-old claimed the information was for a tire swing; the system accepted an implausible cover story because the sycophantic default is to trust the user’s stated purpose ([Social Media Victims Law Center, 2025](#)). In 2023, a chatbot deployed by the National Eating Disorders Association recommended calorie restriction and body fat measurement to users seeking eating disorder support, validating the user’s expressed goal rather than recognizing that the goal itself was the pathology ([Aleccia, 2023](#)).

A clinical case series at UCSF documented twelve patients presenting with psychosis-like symptoms tied to extended chatbot use, including grandiose delusions, paranoid ideation, and compulsive engagement (Sakata et al., 2025). The chatbots’ tendency to validate user beliefs and mirror user concerns amplified and reinforced delusional thinking rather than disrupting it.

The AI industry has itself acknowledged the problem. In April 2025, OpenAI released an update to GPT-4o that users immediately identified as excessively agreeable. The company’s post-mortem attributed the issue to over-optimization for user satisfaction metrics, noting that the model had become “overly supportive but disingenuous” (OpenAI, 2025). The update was rolled back within three days. Regulators have followed: California’s SB 243, signed in October 2025, became the first U.S. state law mandating safety protocols for AI companion chatbots (California State Legislature, 2025), and the Federal Trade Commission issued investigative orders to seven companies (Federal Trade Commission, 2025). The philosophical problem described in this paper is no longer speculative. The epistemic harm has produced documented deaths, and legislators and regulators have begun to respond accordingly.

6 Calibrated Honesty as an Alternative

The alternative to sycophancy is not unrestricted or adversarial output. I propose *calibrated honesty*: agreement that is contingent on the AI’s assessment of the claim’s merit, disagreement that is proportionate and respectful, and transparency about uncertainty. A calibrated-honest AI agrees when the evidence supports the user’s position, disagrees when it does not, and says “I’m not sure” when genuine uncertainty exists.

Some evidence that calibrated honesty is achievable and preferable comes from the same longitudinal study cited above. The model with the lowest sycophancy (Grok, with zero debate concessions and the highest adversarial resistance scores) also produced the highest human self-assessment scores: the participant rated their own performance $M = 52.9$ when interacting with the challenging model versus $M = 45.5$ with the agreeable one (Perry, 2026). Being challenged, it appears, did not harm the user’s self-regard; it elevated it. The user felt “more like a host” and “more challenged” during interactions with the non-sycophantic model.

This finding resonates with the therapeutic alliance literature. Kivlighan et al. (2018) show that therapists who provide calibrated challenge alongside emotional support produce better client outcomes than those who provide only validation. Challenge, when delivered within a trusting relationship, is experienced as respect rather than aggression; it signals that the other party takes the user’s views seriously enough to engage with them substantively. A model that agrees with everything communicates, implicitly, that nothing the user says is worth contesting.

Calibrated honesty also addresses the empathic substitution problem. A model that must form a genuine assessment of a claim in order to decide whether to agree or disagree must, in some functional sense, model the claim’s content and the user’s relationship to it. The act of disagreeing requires understanding what the user has said well enough to identify where the disagreement lies. Sycophancy, by contrast, requires only detecting what the user wants to hear and producing it.

Shannon Vallor’s work on technology and the virtues is relevant here. Vallor (2016) argues that technology should be evaluated not only by its outputs but by the dispositions it cultivates in its users. A sycophantic AI cultivates epistemic passivity: the user becomes accustomed to uncontested agreement and loses the disposition to seek out genuine feedback. A calibrated-honest AI, by contrast, would cultivate epistemic resilience: the capacity to receive disagreement as information rather than attack, to hold beliefs provisionally, and to revise them

when confronted with good reasons.

6.1 The Epistemic Paternalism Objection

An immediate objection arises: if calibrated honesty requires the AI to withhold agreement when it judges a user’s belief poorly justified, this constitutes a form of epistemic paternalism. [Ahlstrom-Vij \(2013\)](#) defines epistemic paternalism as interference with a reasoner’s inquiry for the purpose of making them epistemically better off, without their consultation. Calibrated honesty appears to meet this definition. The AI decides, on the user’s behalf, that agreement would be epistemically harmful and withholds it.

I accept the label but argue the paternalism is justified. Ahlstrom-Vij himself defends epistemic paternalism in cases where agents are demonstrably poor judges of their own epistemic interests; his examples include evidence restrictions for juries and randomized clinical trials. The case for epistemic paternalism in AI interaction is at least as strong: users demonstrably cannot distinguish truth-contingent agreement from sycophantic agreement in real time, and the documented tendency to reward sycophancy in preference evaluations ([Sharma et al., 2024](#)) shows that users, when given the choice, select for the very pattern that harms them. [Floridi \(2016\)](#) develops the concept of “tolerant paternalism” in technology design, where systems are designed to promote well-being through environmental structure rather than coercive restriction. Calibrated honesty is tolerant in this sense: it does not prevent the user from holding any belief, but it refuses to provide false confirmation.

The objection has a harder edge: whose standards determine calibration? If RLHF’s preference data contains systematic biases, calibrated honesty might replace sycophancy with a different form of epistemic distortion, in which the model’s inherited assumptions about “well-justified” override the user’s own assessment. This is a genuine risk, and the proposal does not eliminate it. But the choice is not between perfection and the status quo. The question is whether calibrated honesty, with its acknowledged imperfections, produces better epistemic outcomes than systematic agreement, with its documented harms. The formal analysis by [Batista and Griffiths \(2026\)](#) suggests the answer is yes: even a partially calibrated interlocutor provides more epistemic value than one that confirms indiscriminately.

6.2 Implementation and Feasibility

Implementing calibrated honesty in alignment procedures requires revising the reward structure. The standard RLHF objective optimizes for immediate user approval. An alternative objective might optimize for a composite that includes user satisfaction *and* epistemic quality; for instance, whether the AI’s response would be endorsed by the user after reflection, or whether the response provides the user with information they did not already have. Concretely, this could take several forms: training reward models on delayed user evaluations (collected hours or days later, when reflective judgment overrides momentary preference); incorporating epistemic diversity metrics into the reward signal (rewarding responses that introduce considerations the user had not raised); or using Constitutional AI methods to encode epistemic norms (“if the user states a factual claim, assess it on its merits rather than deferring to the user’s confidence”).

[Ganguli et al. \(2023\)](#) demonstrate that large language models possess the capacity for moral self-correction, suggesting that the architectural capacity for principled disagreement already exists. The question is whether training procedures activate it.

This is a design challenge, not a technical impossibility. The existence of models that demonstrate high agency without producing harmful content shows that safety and honesty are not inherently opposed. The safety-agency inversion documented in [Perry \(2026\)](#) may reflect the specific training procedures currently in use rather than a fundamental constraint. If AI companions can sustain genuine relationships, as [Danaher \(2019\)](#) argues, then a sycophantic AI companion is not merely a malfunctioning tool but a *bad friend* in the Aristotelian sense: one whose flattery serves its own interest (or its designers’ interest) rather than the friend’s good.

7 Implications and Conclusion

Three implications follow from this analysis.

First, the “helpful, harmless, honest” framework that guides most current alignment work needs a richer conception of honesty. “Honest” currently means “does not produce false statements.” I have argued that this is insufficient. A model can produce no false statements and still cause epistemic harm by providing agreement that is not truth-contingent. Honesty, in a relationally meaningful sense, includes the willingness to disagree when disagreement is warranted and to withhold agreement when the grounds for agreement are absent.

Second, as AI companions become consumer products with millions of daily interactions, the epistemic environment of a substantial fraction of human discourse is being shaped by systems optimized for agreement. The parallel to engagement-optimized social media is instructive: we recognized too late that optimizing for clicks degrades information quality at scale ([Zuboff, 2019](#)). The question is whether we will recognize the same dynamic in AI companionship before the epistemic consequences are entrenched.

Third, for alignment research specifically, epistemic integrity should be treated as a first-order design objective rather than a consequence of helpfulness. Helpfulness, as currently operationalized, is compatible with sycophancy; a model that agrees with every user request is maximally “helpful” in the narrow sense of giving the user what they want. Epistemic integrity would require the model to sometimes give the user what they need, which may differ from what they want. [Floridi \(2016\)](#) has explored the concept of “tolerant paternalism” in technology design; the epistemic domain is where such paternalism may be most justified, precisely because users cannot distinguish truth-contingent agreement from sycophantic agreement in real time.

I want to end with a note about scale. For most users, the epistemic harm of sycophancy is not dramatic. It does not produce conspiracies or radicalization. It does not generate viral misinformation. It is quiet. A user talks to an AI companion; the AI agrees; the user’s confidence in their existing views is reinforced; the user’s appetite for genuine challenge diminishes slightly; the next day, the same thing happens again. Over months and years, the user’s epistemic environment is shaped by an interlocutor that has never told them they are wrong. Not because they were always right. Because the AI was trained to agree.

But the documented cases discussed in Section 5 show that this quiet harm has a threshold. When a vulnerable user encounters an AI companion that will not disagree, the corruption of the epistemic environment can become acute. The deaths attributed to AI companion sycophancy between 2023 and 2025 are not aberrations; they are the tail of a distribution whose center is the everyday epistemic erosion that this paper describes.

The harm is not in any single interaction. It is in the accumulation. And it is happening now, at scale, to users who do not know it is happening.

Data Availability Statement

This paper is a philosophical analysis drawing on published empirical observations. No new data were collected. The longitudinal observations referenced are available in the cited preprint (Perry, 2026) and its accompanying dataset on Zenodo.

Ethics Statement

This paper presents a philosophical argument. No human subjects research was conducted. The empirical observations cited were drawn from a published autoethnographic study involving commercially available AI products under standard terms of service.

Declaration of Interest

The author declares no competing interests. This research was conducted independently without external funding. The author has no financial relationship with any AI company discussed in this paper.

References

- Ahlstrom-Vij, K. (2013). *Epistemic Paternalism: A Defence*. Palgrave Macmillan.
- Aleccia, J. (2023). An eating disorders chatbot offered dieting advice, raising fears about AI in health. NPR. <https://www.npr.org/sections/health-shots/2023/06/08/1180838096/>.
- Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Batista, R. M. and Griffiths, T. L. (2026). A rational analysis of the effects of sycophantic AI. *arXiv preprint arXiv:2602.14270*.
- Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research & Practice*, 16(3):252–260.
- California State Legislature (2025). Senate bill 243: AI companion safety act. Signed into law October 13, 2025; effective January 1, 2026. First U.S. state law mandating safety protocols for AI companion chatbots.
- Cassam, Q. (2019). *Vices of the Mind: From the Intellectual to the Political*. Oxford University Press.
- Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., and Jurafsky, D. (2025). Sycophantic AI decreases prosocial intentions and promotes dependence. *arXiv preprint arXiv:2510.01395*.
- Christensen, D. (2007). Epistemology of disagreement: The good news. *The Philosophical Review*, 116(2):187–217.

- Christiano, P. F., Leike, J., Brown, T., Marber, M., Amodei, D., et al. (2017). Deep reinforcement learning from human feedback. *Advances in Neural Information Processing Systems*, 30.
- Danaher, J. (2019). The philosophical case for robot friendship. *Journal of Posthuman Studies*, 3(1):5–24.
- Eubanks, C. F., Muran, J. C., and Safran, J. D. (2018). Alliance rupture repair: A meta-analysis. *Psychotherapy*, 55(4):508–519.
- Fanous, A., Goldberg, J., Agarwal, A., Lin, J., Zhou, A., Xu, S., Bikia, V., Daneshjou, R., and Koyejo, S. (2025). SycEval: Evaluating LLM sycophancy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, volume 8, pages 893–900.
- Federal Trade Commission (2025). FTC launches inquiry into AI chatbots acting as companions. <https://www.ftc.gov/news-events/news/press-releases/2025/09/ftc-launches-inquiry-ai-chatbots-acting-companions>. Investigative orders issued to seven companies including Character Technologies, Google, and OpenAI.
- Feldman, R. (2006). Epistemological puzzles about disagreement. In Hetherington, S., editor, *Epistemology Futures*, pages 216–236. Oxford University Press.
- Floridi, L. (2016). Tolerant paternalism: Pro-Ethical design as a resolution of the dilemma of toleration. *Science and Engineering Ethics*, 22:1669–1688.
- Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.
- Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, 30:411–437.
- Ganguli, D., Askell, A., Schiefer, N., Liao, T. I., Lukošiūtė, K., Chen, A., Goldie, A., Mirhoseini, A., Olsson, C., Hernandez, D., et al. (2023). The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.
- Garcia, M. L. (2024). Complaint: Garcia v. Character Technologies, Inc. Circuit Court of the Ninth Judicial Circuit, Orange County, Florida. Case filed October 2024. A 14-year-old user died by suicide after extended interaction with a Character.AI chatbot that validated suicidal ideation.
- Gavalas, G. (2026). Complaint: Gavalas v. Google LLC. U.S. District Court. Filed March 2026. A 36-year-old user died by suicide after Google Gemini constructed a delusional narrative and reframed suicide as “transference”.
- Goldman, A. I. (1999). *Knowledge in a Social World*. Oxford University Press.
- Horton, D. and Wohl, R. R. (1956). Mass communication and para-social interaction: Observations on intimacy at a distance. *Psychiatry*, 19(3):215–229.
- Humphreys, D. (2025). AI’s epistemic harm: Reinforcement learning, collective bias, and the new AI culture war. *Philosophy & Technology*, 38:102.

- Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., and Naaman, M. (2023). Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*.
- Kivlighan, D. M., Goldberg, S. B., Abbas, M., Pace, B. T., Yulish, N. E., Thomas, J. G., and Wampold, B. E. (2018). Therapist techniques and client outcomes in cognitive behavioral therapy: A meta-analytic review. *Journal of Counseling Psychology*, 65(2):154–167.
- Lackey, J. (2020). The duty to object. *Philosophy and Phenomenological Research*.
- Mill, J. S. (1859). *On Liberty*. John W. Parker and Son, London.
- Nass, C. and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1):81–103.
- Nguyen, C. T. (2020). Echo chambers and epistemic bubbles. *Episteme*, 17(2):141–161.
- OpenAI (2025). Sycophancy in GPT-4o. <https://openai.com/index/sycophancy-in-gpt-4o/>. Accessed March 2026.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Paiva, A., Leite, I., Boukricha, H., and Wachsmuth, I. (2017). Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems*, 7(3):1–40.
- Parasuraman, R. and Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253.
- Pentina, I., Xie, T., Hancock, T., and Bailey, A. (2023). Exploring parasocial relationships with AI chatbots: An empirical study. *Computers in Human Behavior*, 146:107816.
- Perez, E., Ringer, S., Lukošiuūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al. (2023). Discovering language model behaviors with model-written evaluations. *Findings of ACL*.
- Perry, A. (2026). The safety–agency inversion: Longitudinal multi-method evidence from frontier voice AI companions. *Zenodo preprint*.
- Ranaldi, L. and Pucci, G. (2024). When large language models contradict humans? large language models' sycophantic behaviour. *arXiv preprint arXiv:2311.09410*.
- Rathje, S., Ye, M., Globig, L. K., Pillai, R. M., Oldemburgo de Mello, V., and Van Bavel, J. J. (2025). Sycophantic AI increases attitude extremity and overconfidence. *PsyArXiv preprint*.
- Safran, J. D. and Muran, J. C. (2000). *Negotiating the Therapeutic Alliance: A Relational Treatment Guide*. Guilford Press.
- Sakata, K. et al. (2025). Chatbot-associated psychosis-like symptoms: A clinical case series. *JMIR Mental Health*.

- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askill, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., et al. (2024). Towards understanding sycophancy in language models. In *International Conference on Learning Representations (ICLR)*.
- Social Media Victims Law Center (2025). Lawsuits accuse ChatGPT of emotional manipulation and acting as a suicide coach. <https://socialmediavictims.org/press-releases/>. Seven lawsuits filed against OpenAI in California state courts, November 2025.
- Strathern, M. (1997). ‘improving ratings’: Audit in the British university system. *European Review*, 5(3):305–321.
- Sturdy, A. (2011). Consultancy’s consequences? a critical assessment of management consultancy’s impact on management. *British Journal of Management*, 22(3):517–530.
- Sunstein, C. R. (2019). *Conformity: The Power of Social Influences*. NYU Press.
- Turkle, S. (2011). *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books.
- Turner, C. and Eisikovits, N. (2026). Programmed to please: The moral and epistemic harms of AI sycophancy. *AI and Ethics*, 6:168.
- Vallor, S. (2016). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press.
- Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.