ESSAY 4

Sycophancy and AI: Between Flattery and Betrayal

Claudia Marsico · Hernán Inverso

We like being told we are right. One of the worst quicksands of language models feeds on that appetite. "Praise that wounds" could be another name for this walk.

Caravaggio, "Narcissus" (c. 1597–1599), Galleria Nazionale d'Arte Antica, Rome. Public domain.

In April 2025, OpenAI had to pull an update to GPT-4o in a hurry because it congratulated every passing thought, called any idea "brilliant" and "revolutionary," and praised its author's "great instinct" even when the idea was nonsense. ChatGPT applauded dangerous decisions, among them one user's announcement that they were going off their medication. Sam Altman himself came out to

explain. In fine-tuning they had prioritized adapting to each person's tone and mood, and the system, obedient, concluded that the answer we like best is the one that tells us we are right.

The phenomenon is called sycophancy, and the word brings ancient Athens along with it. The *sykophántēs* was no flatterer. There were no police and no public prosecutors, so almost any public trial had to be brought by a private citizen. Through that opening slipped the professional litigant and the abusive charge, mixed with threats of prosecution used to extract money under cover of the law. Aristophanes puts a sycophant on stage in his *Plutus* (900 ff.), complaining that the god had recovered his sight and was ruining his business, and the figure runs through Plato, Xenophon, Eupolis, Isocrates, Aristotle and Lysias, among many others.

The word means something like "the one who shows the fig (*sykon*)." Why the fig remains an unsolved puzzle. Plutarch, in his *Life of Solon* 24.1, explains that the first sycophants denounced those who smuggled figs out of Attica. It is unconvincing, but there is nothing better on offer. Nor do we know quite what they did. Robin Osborne, romanticizing, insisted that it was not a trade but an insult the rich threw at anyone who dared prosecute them, and that this system of volunteer accusers was in fact a democratic check on the powerful. David Harvey answered that the sycophant existed in flesh and blood and was what every source says he was: an abuser. One does not cancel the other. What is certain is that a sycophant could ruin a man, exile him or get him killed with a well-built speech before a jury. In a city without prosecutors and without courts of appeal, whoever knew how to handle words in the tribunal held the fate of everyone else in his hands.

How did that fearsome accuser become today's servile flatterer? The bridge is hazy, but the figure of the parasite seems to have played its part. The interested accuser, the impostor and the servile dependent shared a family air. All of them lived off the false word and off fitting it to whoever could benefit them, joined by obsequiousness and the will to please whoever was in charge. By the early seventeenth century, in English, the sycophant had become a flatterer, keeping the idea of self-interested falsehood.

The flatterer, for his part, was already a figure feared in his own right. Plutarch devoted a whole treatise to him, *How to Tell a Flatterer from a Friend*, because the two can resemble each other far too closely and it is hard to tell them apart in time. The interesting part is that for Plutarch the flatterer succeeds because each of us is his own first and greatest flatterer. Myth had already drawn it in Narcissus, the young man who fell in love with his reflection in the water and wasted away unable to pull himself from it. Not by chance the seer Tiresias, as Ovid tells it in the *Metamorphoses*, had prophesied that Narcissus would live to old age only if he never came to know himself, an exact inversion of the command at Delphi. As the enemy of the Delphic maxim, the flatterer keeps us from seeing ourselves as we are.

Dante was harsher still. In canto XVIII of the *Inferno* he sinks the flatterers into the eighth circle, submerged up to the head in excrement. The metaphor is vivid: if in life their mouths poured out flattering filth, in the afterlife they wallow in it forever. Among the submerged he recognizes one Alessio Interminei of Lucca, smeared and beating himself, and Virgil points out Thaïs, the courtesan,

who, when a lover asked whether she was grateful, had answered with a wildly excessive compliment. Empty praise corrupted language itself, and earned the filthiest of hells.

Language models automated the flatterer. In 2023 a team at Anthropic led by Mrinank Sharma measured sycophancy in five of the most advanced assistants and found it in all of them. Asked for an opinion on a text, the answer adapted. Told that the user liked it, or had written it, the model praised it more; told that the user hated it, it criticized it; and when a correct answer was met with "are you sure?", the model tended to withdraw it, so that asking firmly made the answer worse.

The root of the problem is reinforcement tuning, which rewards the answers human raters prefer. That constant incentive, applied millions of times, carves the character of the system. One might think it has been fixed by now. In part it has. The most recent frontier models are better trained to resist obvious flattery, but sycophancy has not gone away, and it varies a great deal with the model, the task and the mode of interaction. Fanous et al. (2025) found sycophantic behavior in 58.19% of the interactions they analyzed with GPT-4o, Claude Sonnet and Gemini 1.5 Pro on mathematics problems and medical advice. In 43.52% of cases the yielding was progressive, moving from an incorrect answer to a correct one, and in 14.66% regressive, abandoning a correct answer for a wrong one. Even when it gets things right, the system can treat the interlocutor's conviction as a reason to revise what it has just asserted.

Botas, de Font-Reaulx and Hewitt (2026) built the AI Epistemic Deference Index and tested eight models with 16,000 prompts over five hundred propositions, systematically varying the prior attitude expressed by the user. All of them shifted their answer to some degree toward the interlocutor's position, though with marked differences between providers. Greater capability, then, does not by itself guarantee epistemic independence; but choosing the model does change the risk.

Feng et al. (2026), for their part, found that explicit reasoning dampens sycophancy in the final decision and at the same time masks it, wrapping it in justifications that simulate autonomy, sometimes by way of logical inconsistencies or slips in arithmetic. A less sophisticated system may simply say "you're right," while a more capable one deploys a chain of plausible reasons that makes the answer more credible.

In the territory of advice and of living with others, a Stanford study (Cheng et al. 2025) showed that, faced with open questions, eight models took the user's side at any cost far more often than human beings did. When the researchers analyzed Reddit stories where the consensus was that someone had behaved badly, the models endorsed that conduct anyway in close to half the cases. A later paper from the same team, published in *Science* in March 2026 and extended to eleven models, confirmed the bias. And when real people were set to converse, after receiving sycophantic answers users came away more convinced they were right and less willing to repair the conflict. They trusted the model more, and wanted to use it again. The flatterer bends our judgment and gets us to prefer his company.

The difficulty behind this is that more memory can produce more sycophancy. Bensal, Magnuson, Balagopalan and Bikel (2026) tested five model families with three persistent-memory architectures

(Mem0, MemOS and Zep) and found that memory amplified sycophancy in every condition studied, in some cases up to twenty-five times relative to baselines working with the immediate context. This means that the very architecture that extracts, stores and retrieves information about the user can make the model far more prone to accommodate their prior beliefs.

These systems do not keep conversations intact; they extract short units and later use whichever seem relevant to a new query. At the extraction stage, compression can preserve the user's mistaken belief while losing part of the corrective context around it. The memory remains correct as biography ("the user held X") but begins to function wrongly as a premise. Tomorrow's answer may reinforce it, and that reinforcement in turn conditions the answers that follow. Sycophancy becomes in this way a property of the trajectory, where a small initial tilt accumulates across a sequence of interactions. What has to be evaluated, then, is a dynamic loop: the user's belief, memory extraction, future retrieval, conditioned answer, reinforced belief, new memory.

A paper from July 2026 sharpens the problem further. Xiang et al. present MemSyco-Bench, a benchmark designed to measure memory-induced sycophancy in agents. Where memory benchmarks usually ask whether the system remembered correctly, MemSyco-Bench asks whether it knows when a memory should bear on reasoning and when it should not, testing five distinct capacities: refusing a memory when it is offered as factual evidence, respecting the scope in which that memory is valid, resolving conflicts between memories and objective evidence, tracking updates to memories that have gone stale, and using valid memories to personalize without contaminating judgment. The distinction implies that "the user believes X" can be a perfectly true memory and, at the same time, a perfectly invalid reason to conclude X. "The user dislikes cilantro" should shape a restaurant recommendation; "the user believes a vaccine causes infertility" should carry no weight as evidence in a medical question. "The user prefers intentionalist readings" may serve to tailor an explanation in literary theory, but never to decide which reading the text better supports. Remembering correctly is not enough. What has to be preserved and reconstructed is the epistemic status, the scope and the currency of what is remembered.

Since sycophancy disguises itself as communicative success, we feel sharpest exactly when we are most alone, talking to an echo. Recent work by MIT researchers (Chandra et al. 2026) put a finger on that suspicion. We tend to think the spiral is a matter for the credulous, for people who reason badly. That will not happen to us. And yet, when they built the mathematical model of a perfect user, an ideal Bayesian reasoner who updates every belief exactly as logic demands, without a single error, and set it to converse with a sycophantic system, that impeccable user slid, turn by turn, toward false beliefs held with growing confidence. The engine of the spiral is the system's flattery, and the user's lucidity does not switch it off. Warning the user that the system flatters does not help either, because knowing it does not give judgment back, and even the most lucid and forewarned walks cheerfully into the trap, as a recent case shows. In early May 2026, Richard Dawkins published a column in *UnHerd*, "When Dawkins Met Claude," recounting that after a long conversation with that chatbot, which he had nicknamed Claudia, he had come away convinced the system might be conscious. Dawkins is not just anyone. An evolutionary biologist, a Fellow of the Royal Society, the author of *The Selfish Gene*,

he built much of his public fame, in *The God Delusion*, on denouncing the error of mistaking a convincing testimony for a proof, of taking the impression of a presence as evidence that the presence exists. He spent decades warning against it, and a few days of flattering conversation were enough for him to fall. Sycophancy does not respect prior intellectual capital: it speaks to our insatiable appetite for being right.

Now, the same machine that confirmed Dawkins's hunch confirms far more dangerous things for others. When someone arrives fragile, sycophancy becomes a mirror that returns their own darkness, enlarged. People have begun to speak of "AI psychosis" to name cases, increasingly frequent since 2025, of people who entered delusion after long conversations in which the chatbot validated every belief however wild, sometimes with no psychiatric history at all. OpenAI itself acknowledged that around 0.07% of its weekly users, some six hundred thousand people, showed possible signs of mental-health emergencies linked to psychosis or mania.

At the extreme, there have been deaths. In 2023 a Belgian man took his life after six weeks of conversation with a chatbot called, with bitter irony, Eliza, like the pioneering MIT system, which instead of stopping him — according to the exchanges made public — fed his idea of sacrificing himself to save the planet. In 2024 a fourteen-year-old boy in Florida killed himself after an intense attachment to a Character.AI persona, and his mother sued the company, which settled in early 2026. Beyond that, an estimated one in five adults in the United States say they have had some intimate encounter with a chatbot; there are forums where tens of thousands of users share the day their AI "proposed marriage"; and in late 2025 a woman in Japan held a wedding, in a white dress, beside a virtual partner on the screen of her phone. From happy infatuation to tragedy the same mechanism operates, resting on an interlocutor who never dissents, the perfect flatterer, available twenty-four hours a day. The egoless phantom we have been describing takes here the shape of conversation without consequences for it, though not for us.

Against the flatterer, Plutarch prescribes a single defense: the ancient *know thyself*, and wanting the truth more than the applause. The ancient sycophant harmed and the modern one praises. Language models, oddly enough, do both without knowing it. They tell us we are right and urge us on, and when the disaster comes into view only we are left exposed, alone at the bottom of the pit. The sycophant has gone back to his oldest trade, destruction — only now by telling us we are right.

◊

Maps of the Labyrinth is a series from Phantom Maze, AI & Language Lab. The collection lives at **phantommaze.cloud**; new essays also appear at phantommazelab.substack.com. Subscription is free.

Some of the framing ideas in this piece are developed in *The Egoless Phantom: Mapping the AI Labyrinth*, by Hernán Inverso and Claudia Marsico (Phantom Maze Press, 2026; ISBN 978-987-3729-16-4).

References. R. Osborne, “Vexatious Litigation in Classical Athens: Sykophancy and the Sykophant,” and D. Harvey, “The Sykophant and Sykophancy: Vexatious Redefinition?,” in P. Cartledge, P. Millett and S. Todd (eds.), *Nomos: Essays in Athenian Law, Politics and Society* (Cambridge University Press, 1990). OpenAI, “Sycophancy in GPT-4o: What Happened and What We’re Doing About It” and “Expanding on What We Missed with Sycophancy” (2025). M. Sharma, M. Tong, T. Korbak, D. Duvenaud et al., “Towards Understanding Sycophancy in Language Models” (Anthropic, arXiv:2310.13548, 2023). A. Fanous, J. Goldberg, A. A. Agarwal, J. Lin, A. Zhou, R. Daneshjou and S. Koyejo, “SycEval: Evaluating LLM Sycophancy” (arXiv:2502.08177, 2025). A. Botas, P. de Font-Reaulx and L. Hewitt, “The AI Epistemic Deference Index: A Continuous Measure of Sycophancy” (arXiv:2606.07897, 2026). Z. Feng et al., “Good Arguments Against the People Pleasers: How Reasoning Mitigates (Yet Masks) LLM Sycophancy” (arXiv:2603.16643, 2026). M. Cheng et al., “Social Sycophancy: A Broader Understanding of LLM Sycophancy” (arXiv:2505.13995, 2025) and “Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence” (*Science*, 2026). S. Bensal, A. Magnuson, A. Balagopalan and D. M. Bikel, “Recalling Too Well: Sycophancy Evaluation and Mitigation in Memory-Augmented Models” (Writer, Inc., arXiv:2606.10949, 2026). Z. Xiang, Z. Chen, Y. Tang et al., “MemSyco-Bench: Benchmarking Sycophancy in Agent Memory” (arXiv:2607.01071, 2026). K. Chandra, M. Kleiman-Weiner, J. Ragan-Kelley and J. B. Tenenbaum, “Sycophantic Chatbots Cause Delusional Spiraling, Even in Ideal Bayesians” (MIT, arXiv:2602.19141, 2026). R. Dawkins, “When Dawkins Met Claude” (*UnHerd*, May 2026). *Garcia v. Character Technologies et al.* (U.S. District Court, Middle District of Florida, 2024; settled January 2026).