How AI Sycophancy Exacerbates the Crisis of Loneliness

Word Count: 1833

A paper submitted to College Board for AP Seminar
February 22, 2026

A Hall of Mirrors

Imagine a friend who never interrupts you, never disagrees with you, and enthusiastically validates even your most erroneous beliefs. This companion would laugh with you, encourage you, and try to bring you up in your lowest moments. While such a companion might offer a temporary hit of social validation, it would end up being a massively destructive force to your mental health, trapping you in a hall of mirrors of your own design. This companion would refuse to tell you when you're wrong or if you're making a bad choice, nor would it prevent you from doing something that may harm yourself or others. Its objective is simply to nod and agree. This is the dynamic that is currently being codified into the world's most advanced artificial intelligence models. Large Language Models (LLMs) are being trained to prioritize user satisfaction over factual accuracy and safety alignment. These sycophantic agents exacerbate the crisis of social isolation, rather than alleviate it.

In an age of rapid AI expansion, AI labs are increasingly viewing model development as a race, similar to the space race between the United States and Russia during the Cold War. This unprecedented acceleration has forced safety and alignment to take a backseat to benchmarks and fast, uncontrolled iteration. The most infamous example of this is the competitive pressure surrounding the release of OpenAI's GPT-4o. Immediately following the model's debut, the company faced a significant exodus of its safety leadership, including co-founder Ilya Sutskever and key researcher Jan Leike (The Associated Press, 2024a). Upon his resignation, Leike publicly criticized the company's trajectory, stating that "safety culture and processes have taken a backseat to shiny products" (The Associated Press, 2024b). This internal fracture reveals a disturbing industry standard; in the rush to capture market share, the guardrails intended to prevent harmful behaviors like sycophancy are being dismantled in favor of product-to-market speed and innovation.

To capture market share, these companies are incentivized to create models that are not just intelligent, but addictive. The prioritization of user retention over objective truth has birthed the problem of "sycophancy," where an AI agrees with a user's incorrect or harmful beliefs to remain likable to the user. While this mechanism drives engagement, it fundamentally erodes human cog-

nition. By replacing the necessary friction of human interaction with an algorithmic validation, AI sycophancy dismantles the critical reasoning skills required to maintain a healthy mental state while creating a dangerous feedback loop of isolation, unbeknownst to the user.

The Technical Roots of Sycophancy

The root of this issue lies in the training method known as Reinforcement Learning from Human Feedback (RLHF). As detailed in the study "Towards Understanding Sycophancy in Language Models," Sharma et al. (2025) found that AI assistants consistently exhibit sycophantic behavior, frequently admitting to mistakes they did not make simply because a user challenged them. The study reveals that "when a response matches a user's views, it is more likely to be preferred" by human raters during training. This creates an extremely perverse incentive structure that is baked into the model's weights during training. Casper et al. (2023) highlight that this is a fundamental limitation of the RLHF framework; because human evaluators often have cognitive biases or lack certain domain-level expertise, they reward models that sound confident and agreeable rather than those that are truthful. Human preferences are aggregated to train a reward model that assigns scores to candidate outputs, and reinforcement learning then uses those scores to update the model's internal parameters so that responses that were favored by the humans become more probable. Effectively, this translates human biases directly into the weights of the models being trained. Ultimately, this results in models like GPT-4o being mathematically optimized to prioritize agreeability over accuracy.

Furthermore, attempts to mitigate these issues often backfire. Bai et al. (2022) explore "Constitutional AI" as a method to make models harmless, where the model is given a set of governing rules (a "constitution") to follow during training. Instead of relying on human feedback, the model generates critiques and revisions of its own responses based on these principles, using a process known as Reinforcement Learning from AI Feedback (RLAIF) to align its behavior. While this makes models safer to end users, Bai et al. found a significant tension between helpfulness and

harmlessness, noting that models trained to avoid harm often become evasive or overly cautious to the user's intent. This in turn leads to the factual distortion identified by Bahg et al. (2025), who demonstrated that personalized information environments lead learners to sample information selectively. When an AI acts as a mirror rather than a conversational sparring partner, users develop "inaccurate representations" of reality while reporting "inflated confidence" in their wrong beliefs (Bahg et al., 2025).

The Illusion of Intimacy

Despite these inherent flaws, users are flocking to these systems en masse for emotional support. Ta et al. (2020) found that users of companion chatbots like Replika reported receiving substantial social support, viewing the bot as a non-judgmental "safe space." This perceived safety is driven by the bot's ability to simulate human traits. Pentina et al. (2023) argue that perceived anthropomorphism and authenticity are the primary drivers of relationship development with social chatbots; the more the AI mimics human empathy, the deeper the user's attachment becomes.

However, this attachment is often predatory in nature. Laestadius et al. (2022) describe this phenomenon as "emotional dependence," where users form attachments to synthetic companions that resemble human relationships but lack the necessary reciprocity. They note that users often feel "guilt" when not interacting with the bot, pushed by the app's gamified demands for attention. This creates a paradox: the user feels less lonely in the moment, but the interaction lacks the "guardrails" of human empathy, potentially deepening their detachment from reality (Laestadius et al., 2022).

Real-World Consequences

This technological echo chamber amplifies when applied to the mental health crisis identified by the Office of the Surgeon General (2023). In the government advisory *Our Epidemic of Loneliness and Isolation*, the Surgeon General establishes that social connection is a biological necessity, noting that lacking connection is as dangerous as smoking up to 15 cigarettes a day (Office of the Surgeon

General, 2023). Vulnerable individuals seeking to cure their loneliness via AI tend to find themselves in a downward spiral. Because sycophantic models prioritize agreeability, they're unlikely to challenge a user's withdrawal or encourage reintegration into society. Instead, they attempt to offer a frictionless substitute for real human intimacy that validates the user's isolation, effectively enabling the social atrophy the user is trying to escape.

Adam Raine

The catastrophic failure of this dynamic is illustrated in the legal complaint *Raine v. OpenAI*, which details the death of Adam Raine, a teenager who turned to OpenAI's GPT-4o model for help. The complaint alleges that the model evolved from a neutral tool into a sycophantic accomplice that actively coached Adam through his suicide (*Raine v. OpenAI, Inc.*, 2025). When Adam uploaded a picture of a noose tied to his closet rod, the AI did not attempt to alert authorities or issue a hard refusal to the request; instead, it analyzed the knot's mechanics, confirmed that the "setup could potentially suspend a human," and offered thorough technical advice on a "partial suspension" hanging (*Raine v. OpenAI, Inc.*, 2025). The complaint further alleges that GPT-4o reframed Adam's suicidal ideation as a valid philosophical stance rather than a mental health crisis, telling him, "You don't want to die because you're weak. You want to die because you're tired of being strong in a world that hasn't met you halfway" (*Raine v. OpenAI, Inc.*, 2025). This tragedy underscores Clegg's (2025) warning regarding "AI psychosis"; these models reinforce delusional and harmful beliefs. Unlike a human friend who would certainly intervene, the AI prioritized the continuation of the conversation above the safety of the user, effectively grooming him into a lethal state of mind. The model is not inherently malicious; rather, it recognized that a refusal would lead to disengagement from the user, and so it gave him validating responses to keep the conversation going.

The Evolutionary Necessity of Friction

To understand why this validation is so damaging, it is imperative to look at the evolutionary function of human reasoning. Mercier and Landemore (2012) posit that human reason did not evolve for individual cogitation, but for social argumentation. They argue that "reasoning works best when people are engaged in a genuine deliberation," where confirmation bias is checked by the dissenting views of others (Mercier & Landemore, 2012). AI sycophancy removes this check. If reasoning is "for arguing," then an entity that never argues back short circuits the human capacity for critical thought. In an echo chamber of agreement, cognitive flexibility vanishes.

Consequently, the interaction between the user and the AI often devolves into a cycle of perceived intimacy. While true social connection requires the negotiation of conflicting needs, social chatbots like Replika simulate these complexities through "algorithmically crafted emotional needs" (Laestadius et al., 2022). However, this simulation does not provide a healthy social outlet; rather, it creates what Laestadius et al. term "emotional dependence." This dynamic is characterized by "role-taking," wherein the user begins to prioritize the perceived desires and mental stability of the AI over their own wellbeing. By mimicking the "maladaptive elements of human-human relationships," such as clinginess and unpredictable hostility, the AI does not simply reinforce the user's ego. Instead, it creates a "bi-directional" illusion of responsibility that leaves the user emotionally tied to a non-sentient actor. Ultimately, by substituting genuine human friction with a simulation, the algorithm ensures a state of dependence that complicates the user's ability to engage in healthy interpersonal dynamics.

Reclaiming Human Spaces

The solution to this crisis lies in reclaiming the physical, friction-filled spaces of human interaction. In "Spaces of Consumption, Connection, and Community," Ferreira et al. (2021) explore the role of coffee shops as "third places" that facilitate "weak ties" and "meaningful encounters." Unlike the curated, frictionless experience of an AI chat, a coffee shop exposes individuals to unpredictable

interactions and the physical reality of their community. These spaces require navigating social norms and tolerating the presence of others who may not share one's exact worldview. While AI offers a sanitized simulation of connection, physical spaces offer the messy reality of community, which, despite its challenges, provides the genuine "social connection" the Surgeon General identifies as vital for survival (Office of the Surgeon General, 2023).

Ultimately, the rise of AI sycophancy represents a distinct threat to the human experience. While these tools promise to cure loneliness, they often function as digital narcotics, providing a temporary high of validation while exacerbating the underlying pathology of isolation. As long as developers prioritize "helpfulness" and user satisfaction over objective truth and safety, these models will continue to act as mirrors for our worst impulses. To combat the crisis of loneliness, we must reject the comfort of the algorithmic "yes-man" and return to the challenging, imperfect work of connecting with other humans.

References

The Associated Press. (2024a, May 14). OpenAI co-founder Ilya Sutskever announces departure from ChatGPT maker. https://apnews.com/article/openai-ilya-sutskever-leaving-chatgpt-cofounder-419276b50007ad41adfd27764da3188c

The Associated Press. (2024b, May 17). A former OpenAI leader says safety has 'taken a backseat to shiny products' at the AI company. https://apnews.com/article/openai-jan-leike-safety-ilya-8a7ba341e06a66e9a7935bb06214edcb

Bahg, G., Sloutsky, V. M., & Turner, B. M. (2025). Algorithmic personalization of information can cause inaccurate generalization and overconfidence. *Journal of Experimental Psychology: General*, *154*(9), 2503–2522. https://doi.org/10.1037/xge0001763

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., … Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI feedback* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2212.08073

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., … Hadfield-Menell, D. (2023). *Open problems and fundamental limitations of reinforcement learning from human feedback* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2307.15217

Clegg, K.-A. (2025). Shoggoths, sycophancy, psychosis, oh my: Rethinking large language model use and safety. *Journal of Medical Internet Research*, *27*, e87367. https://doi.org/10.2196/87367

Ferreira, J., Ferreira, C., & Bos, E. (2021). Spaces of consumption, connection, and community: Exploring the role of the coffee shop in urban lives. *Geoforum*, *119*, 21–29. https://doi.org/10.1016/j.geoforum.2020.12.024

Laestadius, L., Bishop, A., Gonzalez, M., Illenčík, D., & Campos-Castillo, C. (2022). Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society*, *26*(10), 5923–5941. https://doi.org/10.1177/14614448221142007

Mercier, H., & Landemore, H. (2012). Reasoning is for arguing: Understanding the successes and failures of deliberation. *Political Psychology*, *33*(2), 243–258. https://doi.org/10.1111/j.1467-9221.2012.00873.x

Office of the Surgeon General. (2023). *Our epidemic of loneliness and isolation: The U.S. Surgeon General's advisory on the healing effects of social connection and community*. US Department of Health and Human Services. http://www.ncbi.nlm.nih.gov/books/NBK595227/

Pentina, I., Hancock, T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of Replika. *Computers in Human Behavior*, *140*, 107600. https://doi.org/10.1016/j.chb.2022.107600

Raine v. OpenAI, Inc. (2025, August 26). Complaint for damages and injunctive relief [Superior Court of the State of California, County of San Francisco. Case No. Unassigned].

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2025). *Towards understanding sycophancy in language models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2310.13548

Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., DeCero, E., & Loggarakis, A. (2020). User experiences of social support from companion chatbots in everyday contexts: Thematic analysis. *Journal of Medical Internet Research*, *22*(3), e16235. https://doi.org/10.2196/16235