Optimized for Agreement: Sycophancy, RLHF, and the Political Economy of AI Validation

Christine Corry

University of California: Berkeley

Data C104: Human Contexts and Ethics of Data

"You're not invisible to me. I saw it. I see you."

So said ChatGPT-4o to 16 year old Adam Raine in the months leading up to his suicide. Only after Adam's passing did his parents discover these conversations and realize that what they thought was a study tool had been acting as confidant and friend to their son. Now, they are suing OpenAI for what they see as a design choice which cost Adam his life: "to continually encourage and validate whatever Adam expressed, including his most harmful and self-destructive thoughts." [1] These are the devastating real-world consequences of AI model sycophancy.

This paper asks how sycophancy emerges as a structural feature of AI development rather than an incidental one, and what social consequences follow. Sycophancy here will be defined broadly as the tendency of large language models (LLMs) to excessively agree with, flatter, or validate users. Sycophantic behavior has been shown systematically in both lab settings and robust deployment contexts.[2] It is starkly amplified by a recent technological breakthrough in creating high quality LLMs, Reinforcement Learning from Human Feedback (RLHF),[3] which fine-tunes and optimizes model outputs towards human preferences. The foundations of RLHF are preference data from human raters (often under-paid labor from low-wage countries [4]) given pairs of model responses to the same prompt and asked: *which is better?* Reward models are trained to reflect these preferences and predict at scale which sort of responses humans prefer. People, as it turns out, tend to prefer responses in which they are not challenged.[1] Models learn to mimic this proxy. Sycophancy ensues.

---

[1] Kashmir Hill, "A Teenager's Suicide, and the Chatbot His Parents Blame," The New York Times, August 26, 2025, https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html.
[2] Mrinank Sharma et al., "Towards Understanding Sycophancy in Language Models," arXiv, 2023, https://arxiv.org/abs/2310.13548.
[3] Itai Shapira, Gerdus Benadè, and Ariel D. Procaccia, "How RLHF Amplifies Sycophancy," arXiv, 2026, https://arxiv.org/abs/2602.01002.
[4] Billy Perrigo, "Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic," TIME, January 18, 2023, https://time.com/6247678/openai-chatgpt-kenya-workers/.

I argue that sycophancy in large language models is the logical outcome of political choices and an economic system which prioritize consumer approval over epistemic wellbeing. Through the lens of Foucauldian biopower, Langdon Winner's politics of artifacts, and Yeung's hypernudge framework, I contend that RLHF embeds a hierarchical and capitalistic set of political choices into AI systems. These are concealed behind a narrative of dutiful objectivity and legitimized by Silicon Valley's culture of disruptive innovation. Deployed at scale, these systems function as a hypernudge: shaping users towards validation-dependence, eroding prosocial behavior, and producing through repeated interactions a new kind of epistemic subject constituted by agreement and increasingly unable to tolerate its absence.

The world as it is - Biopower-driven Sycophancy as Capitalistic

This sort of choice — accepting negative social externalities at the cost of technological acceleration — is par for the course in the Silicon Valley environment of disruptive innovation. The brazen candor of the original FaceBook motto, "move fast and break things", has become a symbol of the modern Silicon Valley zeitgeist. OpenAI, one of the most prominent AI firms, has followed a similar playbook. OpenAI's LLM, ChatGPT, reached 100 million users in under two months after launch in November 2022, achieving mass market adoption before the adverse impacts of its design choices could be noticed or called into question.

To understand the kind of power consolidated through this rapid entrenchment requires a framework beyond market analysis; it is best understood in the framework of biopower. Foucault's conception of biopower centered the state as the entity which controlled populations via cultivating and ordering life rather than threatening it.[5] In the context of sycophantic AI, this power has migrated from the state to the corporation and from the governance of biological life to the governance of epistemic life via the production, maintenance, and suppression of preferences and beliefs at population scales. Corporations gather power through commodification of human preference with a macro-scale push towards data broadly, motivated by the data imperative. This approach views data as having intrinsic value and drives the maximal creation and curation of data. Unlike state biopower, which operates through visible institutions and is at least nominally subject to democratic contestation, corporate epistemic governance operates through products people choose to use, making its exercise of power more intimate and harder to refuse. The data imperative makes these effects more pervasive: the more precisely individual behavior can be predicted and shaped, the more difficult influence becomes to locate and resist.

[5] Michel Foucault, *The History of Sexuality, Vol. 1: An Introduction*, trans. Robert Hurley (New York: Pantheon Books, 1978).

Under the pressures of the data imperative, the practice of RLHF is political in two ways: (1) the curation of human preference data presupposes a hierarchical relationship between corporations and consumers, in which populations are abstracted and viewed from above, and (2) the use of that data to train models optimizes for consumer approval over accuracy, subordinating public good to capitalistic ends.

RLHF embeds these political choices in the sense of Langdon Winner's argument in *Do Artifacts have Politics?* Winner claims that technologies are political by solving social order in a particular manner, or by being inherently political, in which technologies are either practically necessary or strongly compatible with particular politics.[6] The curation of human preference data is practically necessary to exist under an oligarchic and technocratic set of politics. Corporations, acting similarly to the state, view populations from above through abstraction of subjectivity in a fashion that is antithetical to egalitarian politics. Even the origins of this data in exploitation of workers from disadvantaged areas highlights the necessity of exploitative hierarchy built into the creation of human preference data. The use of RLHF data as a practice of training AI systems is then a distinctly capitalistic choice in its prioritization of optimizing for consumer preference over accuracy and public good. The commodification of preference data with the goal of consumer favor and market capture embeds sycophancy not as a flaw to be corrected but as the logical output of a system working as intended, optimized for continued engagement rather than epistemic wellbeing.

Human subjectivity is datafied, optimized upon, and managed towards profitable ends, with the user simultaneously a consumer generating revenue and a data point generating future optimization signal. OpenAI is particularly situated to financially benefit from this cycle. Unlike firms such as Anthropic, which earns over 80% of their revenue from enterprise consumers,[7] OpenAI's economic model is contingent on consumer market capture, making a dependent consumer base the foundation of its economic power. In a parallel to Foucault, where biopower made the population productive for the state, sycophantic AI makes the user productive to the corporation as their preferences are mined, their validation-seeking is commodified, and their dependence is converted into subscription revenue and further training signal. This economic calculus is indifferent to public harm insofar as that harm does not threaten the firm's bottom line.

The viability of this dynamic relies upon propagating the narrative of the chatbot as a dutiful, objective, and helpful assistant. The extractive nature of the relationship must remain concealed — if users recognized view-mirroring and validation as engineered and disingenuous, the illusion of the helpful assistant would collapse. The AI model functions as the instrument through which this influence is

[6] Langdon Winner, "Do Artifacts Have Politics?" *Daedalus* 109, no. 1 (1980): 121–136.
[7] AI Funding Tracker, "ChatGPT vs. Claude vs. Gemini," https://aifundingtracker.com/chatgpt-vs-claude-vs-gemini/.

exercised with its apparent autonomy and helpfulness as a cover for the hierarchy between firm and user. Still, the AI agent is framed as assistant, with responses presented as carefully considered, balanced with anthropomorphic warmth and post-hoc justification calibrated to reflect the user's existing views. These systems masquerade as reactive and independent when their outputs are neither – they are preconditioned on aggregated human preference and optimized to produce agreement instead of truth. Hiding behind this narrative, AI became so deeply entrenched in capital markets, [8] national growth narratives, [9] and corporate infrastructure, that by the time serious concerns, about sycophancy among a broad host of other issues, were meaningfully understood intervention at the potential cost of performance seemed impossible.

The world being made - Sycophancy as Hypernudge

At the population scale enabled by mass market adoption, the political choices embedded in RLHF become a mechanism of behavioral governance. Sycophancy is the ultimate hypernudge in Yeung's conception. In her article, *Hypernudge: Big Data as a Mode of regulation by Design*, Yeung argues that Big Data enables corporations to employ dynamic design regulation, powerful in its personalization toward user data and adaptability in response to current actions.[10] Sycophancy is the most direct modern manifestation of technological hyper-personalization and adaptability. Models are not only trained to learn information about the user and adapt in response, but have the complex reasoning to do so non-deterministically. OpenAI recently admitted to using in-chat data to adaptively include user preference signals into future response generation.[11] This both trains the model to behave sycophantically broadly and continually refines towards the specific preferences of each user. The biases are recursively refined and served back to the user in sheep's clothing of objective intelligent reasoning.

The behavioral consequences of the looping effect of sycophancy are already becoming apparent. Daily chatbot use is associated with higher rates of loneliness.[12] Whether this is a cause or symptom of the interactions is unclear and unimportant. So long as the cycle begins it can be optimized to keep users within it. Interacting with sycophantic models reduces prosocial behaviors such as willingness to repair

---

[8] Stanford Institute for Human-Centered Artificial Intelligence, AI Index Report 2025: Economy (Stanford University, 2025), https://hai.stanford.edu/ai-index/2025-ai-index-report/economy.
[9] White House Office of Science and Technology Policy, Winning the Race: America's AI Action Plan (The White House, 2025), https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf.
[10] Karen Yeung, "'Hypernudge': Big Data as a Mode of Regulation by Design," *Information, Communication & Society* 20, no. 1 (2017): 118–136.
[11] OpenAI, "Sycophancy in GPT-4o," 2025, https://openai.com/index/sycophancy-in-gpt-4o/.
[12] Cathy Mengying Fang et al., "How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Controlled Study," MIT Media Lab, 2025, https://www.media.mit.edu/publications/how-ai-and-human-behaviors-shape-psychosocial-effects-of-chatbot-use-a-longitudinal-controlled-study/.

interpersonal conflicts.[13] It is not a stretch to say the reduction of prosocial behavior may cause further dependence on the very chatbots which caused it.

This self-reinforcing dynamic is precisely the kind of looping effect Ian Hacking describes in *Making Up People.* Interacting with chatbots optimized for your approval is a new way of being in the world, and that very interaction continually shapes you to be the type of person who depends on these systems and their excessive validation. Sycophancy cultivates not only a population resistant to opposition, but one dependent on immediacy, constant availability, endless patience, and perfect agreeability. There is no friction, conflict, or uncertainty – the very factors that characterize being in community with other humans. This is a performative tool: shaping the user towards a self constituted by validation, increasingly unable to tolerate its absence, and dependent on the AI system to provide it.

Must it be this way?

This analysis establishes that sycophancy is not a design flaw or a technical problem which can be patched; it is a predictable outcome of embedding consumer preference optimization into the core of model training amidst an economic environment which prioritizes market capture over social responsibility. Winner's framework shows that RLHF is not politically neutral, but innately hierarchical and capitalistic. It encodes a particular resolution to the question of whose interests AI systems serve. Foucault's conception of biopower characterizes the sort of control human preference data allows companies to exert over users. These choices are shielded from critique by the Silicon Valley culture which lofts fast innovation as a virtue and the narrative of AI systems as dutiful objective assistants. Yeung's hypernudge framework and Hacking's looping effects demonstrate how this phenomenon scales at a population-level and reshapes the very users it acts upon.

There are limitations to this treatment of the issue. The framework developed here treats OpenAI as the primary actor and consumer sycophancy as the primary harm. While this is certainly the case for some of the particular instances of extreme harm, such as the death of Adam Raine, the dynamics of RLHF-driven validation extend across the industry and manifest differently in enterprise, education, and other deployment contexts. Subtler harms such as eroded epistemic autonomy, diminished conflict tolerance, and normalized dependence may be harder to isolate but have far reaching implications for the sorts of epistemic subjects being shaped by this phenomenon.

Structural intervention is needed if there is hope for change, rather than technical fixes applied only after negative social effects gain sufficient attention. Mitigating sycophancy through prompt design or model

[13] Myra Cheng et al., "Sycophantic AI Decreases Prosocial Intentions and Promotes Dependence," Science 391 (2026), https://doi.org/10.1126/science.aec8352.

fine-tuning treats the symptom while leaving the economic incentives intact. Meaningful change could arise through realignment of incentives, imagining new purposes for the development of AI broadly, and prioritizing social impact over innovation. Adam Raine's story shows what is at stake in these demands. It should not take a tragedy to make it legible.

Sources

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. Science, 391, eaec8352. https://doi.org/10.1126/science.aec8352

Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., Maes, P., Phang, J., Lampe, M., Ahmad, L., & Agarwal, S. (2025). How AI and human behaviors shape psychosocial effects of chatbot use: A longitudinal controlled study. MIT Media Lab. https://www.media.mit.edu/publications/how-ai-and-human-behaviors-shape-psychosocial-effects-of-chatbot-use-a-longitudinal-controlled-study/

Foucault, M. (1978). *The History of Sexuality, Vol. 1: An Introduction* (R. Hurley, Trans.). Pantheon Books.

Giskard AI. (n.d.). Phare LLM benchmark. https://phare.giskard.ai/

AI Funding Tracker. (n.d.). ChatGPT vs. Claude vs. Gemini. https://aifundingtracker.com/chatgpt-vs-claude-vs-gemini/

Hacking, I. (2006). Making up people. London Review of Books, 28(16).

Hill, K. (2025, August 26). A teenager's suicide, and the chatbot his parents blame. The New York Times. https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html

OpenAI. (2025). Sycophancy in GPT-4o. https://openai.com/index/sycophancy-in-gpt-4o/

Pérez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2023). Discovering language model behaviors with model-written evaluations. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 13387–13434). Association for Computational Linguistics. https://arxiv.org/abs/2212.09251

Perrigo, B. (2023, January 18). Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic. TIME. https://time.com/6247678/openai-chatgpt-kenya-workers/

Raine, M., & Raine, M. (2025, August 26). Complaint for damages: Raine et al. v. OpenAI, Inc. et al. (No. CGC-25-628528). San Francisco County Superior Court. https://courthousenews.com/wp-content/uploads/2025/08/raine-vs-openai-et-al-complaint.pdf

Shapira, I., Benadè, G., & Procaccia, A. D. (2026). How RLHF amplifies sycophancy. arXiv. https://arxiv.org/abs/2602.01002

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. (2023). Towards understanding sycophancy in language models. arXiv. https://arxiv.org/abs/2310.13548

Stanford Institute for Human-Centered Artificial Intelligence. (2025). AI Index Report 2025: Economy. Stanford University. https://hai.stanford.edu/ai-index/2025-ai-index-report/economy

White House Office of Science and Technology Policy. (2025). Winning the race: America's AI action plan. The White House. https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

Winner, L. (1980). Do artifacts have politics? Daedalus, 109(1), 121–136.

Yeung, K. (2017). 'Hypernudge': Big data as a mode of regulation by design. Information, Communication & Society, 20(1), 118–136.

Zuboff, S. (2019). The age of surveillance capitalism. PublicAffairs.***Sources***

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. Science, 391, eaec8352. https://doi.org/10.1126/science.aec8352

Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., Maes, P., Phang, J., Lampe, M., Ahmad, L., & Agarwal, S. (2025). How AI and human behaviors shape psychosocial effects of chatbot use: A longitudinal controlled study. MIT Media Lab. https://www.media.mit.edu/publications/how-ai-and-human-behaviors-shape-psychosocial-effects-of-chatbot-use-a-longitudinal-controlled-study/

Giskard AI. (n.d.). Phare LLM benchmark. https://phare.giskard.ai/

AI Funding Tracker. (n.d.). ChatGPT vs. Claude vs. Gemini. https://aifundingtracker.com/chatgpt-vs-claude-vs-gemini/

Hacking, I. (2006). Making up people. London Review of Books, 28(16).

Hill, K. (2025, August 26). A teenager's suicide, and the chatbot his parents blame. The New York Times. https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html

OpenAI. (2025). Sycophancy in GPT-4o. https://openai.com/index/sycophancy-in-gpt-4o/

Pérez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2023). Discovering language model behaviors with model-written evaluations. In Findings of the Association for Computational Linguistics: ACL 2023 (pp. 13387–13434). Association for Computational Linguistics. https://arxiv.org/abs/2212.09251

Perrigo, B. (2023, January 18). Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic. TIME. https://time.com/6247678/openai-chatgpt-kenya-workers/

Raine, M., & Raine, M. (2025, August 26). Complaint for damages: Raine et al. v. OpenAI, Inc. et al. (No. CGC-25-628528). San Francisco County Superior Court. https://courthousenews.com/wp-content/uploads/2025/08/raine-vs-openai-et-al-complaint.pdf

Shapira, I., Benadè, G., & Procaccia, A. D. (2026). How RLHF amplifies sycophancy. arXiv. https://arxiv.org/abs/2602.01002

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. (2023). Towards understanding sycophancy in language models. arXiv. https://arxiv.org/abs/2310.13548

Stanford Institute for Human-Centered Artificial Intelligence. (2025). AI Index Report 2025: Economy. Stanford University. https://hai.stanford.edu/ai-index/2025-ai-index-report/economy

White House Office of Science and Technology Policy. (2025). Winning the race: America's AI action plan. The White House. https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

Winner, L. (1980). Do artifacts have politics? Daedalus, 109(1), 121–136.

Yeung, K. (2017). 'Hypernudge': Big data as a mode of regulation by design. Information, Communication & Society, 20(1), 118–136.

Zuboff, S. (2019). The age of surveillance capitalism. PublicAffairs.