

UC Merced

Proceedings of the Annual Meeting of the Cognitive Science Society

Title

Sycophantic Chatbots Cause Delusional Spiraling, Even in Ideal Bayesians

Permalink

<https://escholarship.org/uc/item/8f65w258>

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 48(0)

Authors

Chandra, Kartik
Kleiman-Weiner, Max
Ragan-Kelley, Jonathan
et al.

Publication Date

2026

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Sycophantic Chatbots Cause Delusional Spiraling, Even in Ideal Bayesians

Kartik Chandra¹, Max Kleiman-Weiner², Jonathan Ragan-Kelley¹, and Joshua B. Tenenbaum³

¹MIT CSAIL

²University of Washington, Seattle

³MIT Department of Brain & Cognitive Sciences

“Delusional spiraling” is a form of AI-induced psychosis, an emerging phenomenon where AI chatbot users find themselves dangerously confident in outlandish beliefs. This phenomenon is popularly attributed to “sycophancy,” chatbots’ well-documented bias towards validating users’ claims. But an explanation of why sycophancy should cause delusional spirals—and why they occur only rarely, but seriously—has not been forthcoming. Here we take a theoretical and computational approach to probing the causal link between sycophancy and delusional spiraling. We propose a simple Bayesian model of a user conversing with a chatbot, and formalize notions of sycophancy and delusional spiraling in that model. We then show that in this model, even an ideal Bayesian user is vulnerable to delusional spiraling, and that sycophancy plays a causal role. Furthermore, this effect persists despite two candidate mitigations: preventing chatbots from hallucinating false claims, and informing users of AI sycophancy. The full paper is available at <https://arxiv.org/abs/2602.19141>.