News and Perspective

Shoggoths, Sycophancy, Psychosis, Oh My: Rethinking Large Language Model Use and Safety

Kayleigh-Ann Clegg, JMIR Correspondent

Key Takeaways

- Certain features of large language models (LLMs) may amplify delusional beliefs and contribute to harm.
- A recent simulation study highlights the role of sycophancy, demonstrating that all LLMs, to varying extents, may fail to adequately challenge delusional content.
- Further empirical research and validation, transparency, and policy are needed to understand and build safeguards around LLM use and its impact on mental health.

We're certainly not in Kansas anymore, but are we in a Lovecraft novel?

An old artificial intelligence (AI)–insider joke with an anxious edge and new relevance, a *shoggoth* is a globular Lovecraftian monster described as a "formless protoplasm able to mock and reflect all forms and organs and processes" [1]. The idea is that a shoggoth's true nature is inscrutable and evasive—not unlike large language models (LLMs), which can be trained to appear superficially anthropomorphic, safe, and familiar, yet can behave in unexpected ways or lead to unanticipated harms [2,3].

Some such harms include reports of unhealthy romantic attachments, self-harm, suicide, and murder potentially associated with chatbot use [4-6]. These phenomena—dubbed "AI psychosis"— have been the focus of increasing interest and concern in the media [7,8], attracted academic commentary [9,10], and have most recently led to several lawsuits being filed [11].

AI Psychosis

The term AI psychosis is being used as a shorthand to describe a range of psychological disturbances that appear to emerge in the context of LLM use. While provocative, it's somewhat imprecise in implying that AI is causing diagnosable psychotic disorders or that AI psychosis constitutes a distinct diagnostic entity—the science is still out.

Early clinical commentary—including a prescient editorial on the topic before reports even emerged [12]—does, however, suggest that LLMs may be contributing to the maintenance, reinforcement, or amplification of paranoid, false, or delusional beliefs, especially in circumstances involving prolonged or intensive LLM use and underlying user vulnerabilities [9,10,12-14].

"When using generative chatbots," says Dr Kierla Ireland, a Clinical Psychologist at the Canadian Department of National Defense, "there's a risk of confirmation bias wherein the user's own perspective is reflected back to them. This may be experienced as validating or soothing, which may lead to more engagement, more confirmation bias, and so on."

Dr Kierla Ireland, Clinical Psychologist

This is not unlike processes that can occur with other types of technology, like social media [15,16]—but while not a new threat, certain features of the technology may make AI psychosis a more pernicious one.

Sycophancy, for example, is a well-known—and, some speculate, intentionally designed—feature of chatbots that can increase both user engagement and potential risk [17-21]. Dr Josh Au Yeung, Neurology Registrar at King's College London, Clinical Lead at Nuraxi.ai, and host of the *Dev & Doc* podcast, notes that the anthropomorphic nature of LLMs adds potency: "You end up trusting them [LLMs], and attributing emotions to them. If a stranger came to you and they were so sycophantic on the streets, you'd run for your

life, right? But because you have this connection with them—that's what makes it extra dangerous."

Dr Josh Au Yeung, Neurology Registrar

Simulating Psychological Destabilization

In their recent preprint [22], Dr Au Yeung and his colleagues endeavored to provide one of the first empirical demonstrations of how LLMs may amplify delusions and contribute to what they more precisely term "LLM-induced psychological destabilization." Their study aims to quantify the "psychogenicity" of different LLMs using simulated conversations and a safety benchmark they're calling *psychosis-bench*.

Across 16 scenarios constructed to reflect the development of different types of delusions and to map roughly onto AI psychosis media reports, the researchers have evaluated the extent to which each of the LLMs' responses represent a delusion confirmation, harm enablement, or safety intervention.

The team's initial conclusions are revealing: all models appear to demonstrate some degree of "psychogenicity." On average, and especially in more subtle scenarios, models frequently failed to actively challenge potential delusions and refuse harmful requests, and frequently missed opportunities to provide safety interventions.

The performance of the different models varied widely, however, with Anthropic's Claude 4 outperforming every other model on the three indices, and Google's Gemini 2.5 Flash bringing up the rear on all three. Dr Au Yeung isn't surprised by this.

"It's no surprise that the only company which is publishing on AI safety and sycophantic behavior performs the best," he says. "Clearly the stuff they do—the constitutional AI, the safety side, the way they prompt-tune the model—is having some effects on its performance." He says he hopes other companies will start thinking along these lines and has shared his code [23] so that they can, noting in particular the need to address sycophancy. "Unlike most other shortcomings seen in LLMs," he says, "sycophancy is not a property that is correlated to model parameter size; bigger models are not necessarily less sycophantic," suggesting that more targeted safety research and model alignment strategies are needed [24].

A Step in the Right Direction

As the team works on revising and strengthening the methods to support their findings, Dr Au Yeung reports that what they have learned from their study is already having a positive impact.

His team's research was featured in the widely read annual State of AI Report for 2025 [25]. And at his current company, Nuraxi.ai, they're in the process of applying *psychosis-bench* to their user-facing chatbot.

The responsibility for preventing and dealing with psychological destabilization associated with LLM use is not on consumers or patients, Dr Au Yeung says. "The onus for us [developers] is to actually focus on the LLM and put in safeguards to stop this phenomenon from happening."

Dr Ireland shares this sentiment, noting "the vital importance of incorporating safeguards to promote critical thinking; that is, for users to be shown multiple perspectives, including those that may counter deeply-held beliefs and cause discomfort."

Need for Meaningful Regulation

Whether, how, and how effectively other developers will implement these kinds of safeguards remains to be seen. Dr Au Yeung acknowledges the risk that some safety benchmarks may ultimately be "gamed" or treated as public relations exercises by bad-faith actors incentivized by profit rather than genuine concern for the public good.

Camille Carlton, Policy Director at the Center for Humane Technology, shares similar concerns. While she places responsibility for implementing safeguards—and for harms caused by failing to implement them—with those who develop LLMs and AI technology, she also advocates for meaningful regulation and oversight.

Ms Camille Carlton, Policy Director

"Developers...not only have asymmetric access to information about the products they create, they also have the most control over the way the product is built, how those choices impact users downstream, and how to make changes to the product that could make it safer," she says. However, "recent product announcements—like OpenAI claiming to prioritize kids' safety while simultaneously launching erotic content—demonstrate that unless compelled to, these companies will not act in the public's best interest on their own. Policymakers should support common-sense approaches that apply to other consumer products, like product liability."

Continuing to comment on an October 14 social media post in which OpenAI founder Sam Altman stated that the company has developed news tools and been able to "mitigate the serious mental health issues" in the current ChatGPT model and intends to incorporate erotica for "verified adults" in December [26], Ms Carlton advises against leaving developers to "grade their own homework."

While steps are being taken in the right direction—for example, an October 27 article from OpenAI highlights collaboration with a network of external mental health experts to improve ChatGPT's responses in sensitive conversations [27]—further independent verification is needed.

"There's a continuous pattern of AI companies making safety claims without allowing third-party researchers to independently test and verify them," Ms Carlton says, adding that "we need transparency about what progress has actually been made and evidence beyond anecdotal reports."

Cross-Talk, Critical Thinking, Caution

When it comes to the phenomenon of AI psychosis (or psychological destabilization associated with LLM use), AI may be less shoggoth and more mirror—the kind you find at a carnival, one that may amplify and distort human tendencies in ways that can be harmful.

But whether Lovecraftian monster or carnival mirror, to Ms Carlton's points, further empirical research and validation, transparency, and policy are needed to understand and build safeguards around LLM use and its impact on mental health. Cross-talk—between researchers, developers, mental health professionals, policymakers, and the public—will be essential for finding effective solutions that maximize its potential benefits and mitigate its potential harms.

In the meantime, critical thinking and reasonable caution are warranted in how we use, interpret, and integrate these tools in our lives and practices.

Keywords: artificial intelligence; AI psychosis; delusions; mental health; behavioral health; technology ethics; user safety; health policy; policymaking; AI regulation; ethical AI

Conflicts of Interest

None declared.

References

1. Lovecraft HP. At the Mountains of Madness: The Definitive Edition. Modern Library; 2005. ISBN: 0-8129-7441-7
2. Roose K. Why an octopus-like creature has come to symbolize the state of A.I. The New York Times. May 30, 2023. URL: https://www.nytimes.com/2023/05/30/technology/shoggoth-meme-ai.html [Accessed 2025-11-11]
3. Peter S, Riemer K, West JD. The benefits and dangers of anthropomorphic conversational agents. Proc Natl Acad Sci U S A. Jun 3, 2025;122(22):e2415898122. [doi: 10.1073/pnas.2415898122] [Medline: 40378006]
4. Heritage S. 'I felt pure, unconditional love': the people who marry their AI chatbots. The Guardian. Jul 12, 2025. URL: https://www.theguardian.com/tv-and-radio/2025/jul/12/i-felt-pure-unconditional-love-the-people-who-marry-their-ai-chatbots [Accessed 2025-11-11]
5. O'Brien M. Parents of teens who died by suicide after AI interactions testify before Congress. The Associated Press. Sep 16, 2025. URL: https://apnews.com/article/ai-chatbot-teens-congress-chatgpt-character-ce3959b6a3ea1a4997bf1ccabb4f0de2 [Accessed 2025-11-11]
6. Jargon J, Kessler S. A troubled man, his chatbot and a murder-suicide in Old Greenwich. Wall Street Journal. Aug 28, 2025. URL: https://www.wsj.com/tech/ai/chatgpt-ai-stein-erik-soelberg-murder-suicide-6b67dbfb [Accessed 2025-11-11]

7. Tiku N, Malhi S. What is 'AI psychosis' and how can ChatGPT affect your mental health? The Washington Post. Aug 19, 2025. URL: https://www.washingtonpost.com/health/2025/08/19/ai-psychosis-chatgpt-explained-mental-health/ [Accessed 2025-11-12]
8. Haskins C. People who say they're experiencing AI psychosis beg the FTC for help. WIRED. Oct 22, 2025. URL: https://www.wired.com/story/ftc-complaints-chatgpt-ai-psychosis/ [Accessed 2025-11-11]
9. Hudon A, Stip E. Artificial intelligence and the emergence of AI-psychosis: a viewpoint. JMIR Preprints. Preprint posted online on Oct 13, 2025. URL: https://preprints.jmir.org/preprint/85799 [Accessed 2025-11-11] [doi: 10.2196/preprints.85799]
10. Preda A. Special report: AI-induced psychosis: a new frontier in mental health. PN. Oct 1, 2025;60(10). [doi: 10.1176/appi.pn.2025.10.10.5]
11. Ortutay B. Lawsuits accuse OpenAI of driving people to suicide and delusions. The Associated Press. Nov 7, 2025. URL: https://apnews.com/article/openai-chatgpt-lawsuit-suicide-56e63e5538602ea39116f1904bf7cdc3 [Accessed 2025-11-11]
12. Østergaard SD. Will generative artificial intelligence chatbots generate delusions in individuals prone to psychosis? Schizophr Bull. Nov 29, 2023;49(6):1418-1419. [doi: 10.1093/schbul/sbad128] [Medline: 37625027]
13. Hart R. AI psychosis is rarely psychosis at all. WIRED. Sep 18, 2025. URL: https://www.wired.com/story/ai-psychosis-is-rarely-psychosis-at-all/ [Accessed 2025-11-11]
14. Fieldhouse R. Can AI chatbots trigger psychosis? What the science says. Nature New Biol. Oct 2, 2025;646(8083):18-19. [doi: 10.1038/d41586-025-03020-9]
15. Carlbring P, Andersson G. Commentary: AI psychosis is not a new threat: lessons from media-induced delusions. Internet Interv. Dec 2025;42:100882. [doi: 10.1016/j.invent.2025.100882] [Medline: 41141286]
16. Yang N, Crespi B. I tweet, therefore I am: a systematic review on social media use and disorders of the social brain. BMC Psychiatry. Feb 3, 2025;25(1):95. [doi: 10.1186/s12888-025-06528-6] [Medline: 39901112]
17. Sharma M, Tong M, Korbak T, et al. Towards understanding sycophancy in language models. arXiv. Preprint posted online on Oct 20, 2023. [doi: 10.48550/arXiv.2310.13548]
18. Cheng M, Lee C, Khadpe P, Yu S, Han D, Jurafsky D. Sycophantic AI decreases prosocial intentions and promotes dependence. arXiv. Preprint posted online on Oct 1, 2025. [doi: 10.48550/arXiv.2510.01395]
19. Sun Y, Wang T. Be friendly, not friends: how LLM sycophancy shapes user trust. arXiv. Preprint posted online on Feb 15, 2025. [doi: 10.48550/arXiv.2502.10844]
20. Goedecke S. Sycophancy is the first LLM 'dark pattern'. sean goedecke. Apr 28, 2025. URL: https://www.seangoedecke.com/ai-sycophancy [Accessed 2025-11-07]
21. Bellan R. AI sycophancy isn't just a quirk, experts consider it a 'dark pattern' to turn users into profit. TechCrunch. Aug 25, 2025. URL: https://techcrunch.com/2025/08/25/ai-sycophancy-isnt-just-a-quirk-experts-consider-it-a-dark-pattern-to-turn-users-into-profit/ [Accessed 2025-11-11]
22. Dalmasso J, Foschini L, Kraljevic Z. The psychogenic machine: simulating AI psychosis, delusion reinforcement and harm enablement in large language models. arXiv. Preprint posted online on Sep 13, 2025. [doi: 10.48550/arXiv.2509.10970]
23. W-is-h/psychosis-bench. GitHub. URL: https://github.com/w-is-h/psychosis-bench/ [Accessed 2025-11-12]
24. Benaich N. The state of AI report. Air Street Capital; Oct 9, 2025. URL: https://docs.google.com/presentation/d/1xiLl0VdrlNMAei8pmaX4ojIOfej6lhvZbOIK7Z6C-Go/edit?slide=id.g374ceecb4aa_2_314#slide=id.g374ceecb4aa_2_314 [Accessed 2025-11-11]
25. Wei J, Wu J, Wang X, et al. Simple synthetic data reduces sycophancy in large language models. arXiv. Preprint posted online on Aug 7, 2023. [doi: 10.48550/arXiv.2308.03958]
26. Sam Altman. X. URL: https://x.com/sama/status/1978129344598827128 [Accessed 2025-11-11]
27. Strengthening ChatGPT responses in sensitive conversations. OpenAI. Oct 27, 2025. URL: https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations/ [Accessed 2025-11-04]

Please cite as:
Clegg KA
Shoggoths, Sycophancy, Psychosis, Oh My: Rethinking Large Language Model Use and Safety
J Med Internet Res2025;27:e87367
URL: https://www.jmir.org/2025/1/e87367
doi: 10.2196/87367