

Conversational AI and psychosis: A technological folie à deux

Jinghui Huang^{1,2} MRCPsych, Douglas Yeung Leong^{1,2} MRCPsych, Siow Ann Chong^{2,3} FAMS, Mythily Subramaniam^{2,3} PhD

ABSTRACT

Generative artificial intelligence (AI) is a recent technological advancement that has become increasingly ubiquitous in its usage. Conversational AI, powered by large language models (LLMs) in particular, has opened many possibilities in psychiatry for providing psychoeducation, mental health support and even evidence-based therapy for people with mental health concerns. Despite its possible benefits, there are inherent limitations and dangers to this technology. In this commentary, the authors explore the possible link between psychosis and conversational AI with the use of a case illustration. The authors show how the sycophantic and excessively agreeable nature of LLMs, combined with its use of human-like language and capacity to generate inaccurate information, might collude with and reinforce delusional beliefs of a user with psychosis. Further research on the interaction between conversational AI and psychopathology would be essential in the process of incorporating such a powerful tool into clinical practice, and would aid in developing the necessary guidelines and safeguards to optimise its inevitable use, and minimise its abuse and untoward effects.

Keywords: AI, artificial intelligence, chatbot, psychosis, psychiatry, mental health

INTRODUCTION

Generative artificial intelligence (AI) has shown considerable promise in many fields of medicine. The use of machine learning algorithms and their ability to learn from vast datasets has helped synthesise clinical information, predict outcomes and make diagnoses in a wide range of medical fields.¹ Large language models (LLMs) are a type of generative AI that utilises datasets and deep learning to comprehend and respond to human language prompts. ChatGPT, a conversational AI chatbot powered by an LLM and released by OpenAI in 2022, had reached 100 million active users just 2 months after its launch, making it the fastest-growing application in history.²

LLMs have a wide range of applications in medicine. They are able to summarise information quickly, such as in the creation of discharge summaries, and also have uses in medical education.³ There has also been a recent trend of users turning towards LLMs as therapists. LLMs, with their linguistic capabilities, show potential in identifying and reframing unhelpful thoughts, similar to how a therapist might in psychological therapies such as cognitive behavioural therapy.⁴ They are “sensitive” to what users express and can provide validating statements to users. The capacity of LLMs for emotional recognition and adaptability to user input provides compelling evidence for their application in addressing social isolation.⁵ They are also scalable and available at all times unlike traditional methods that are limited by manpower and waking hours.

Despite the potential benefits, there are also inherent dangers regarding the use of this technology for mental health support. Therapists often employ techniques, such as therapeutic confrontation to highlight incongruencies in the words and behaviours of patients, to encourage self-reflection and help reframe negative cognitions. These tasks may not be performed adequately by conversational AI, especially if there is no explicit prompt to do so. Furthermore, conversational AI such as ChatGPT are designed to be supportive, and when done excessively could run the risk of sounding sycophantic.⁶ Many developers of conversation AI applications recognise this and have implemented guardrails to monitor for potentially dangerous inputs (such as requests to assist with crime and suicide), and to modify outputs to conform to ethical guidelines.⁷ OpenAI has shared that they have been training models not to provide self-harm instructions and to provide direction to seek help in the real world, although they acknowledge that there are moments when their model might not always do so.⁸ These guardrails might face greater difficulties in identifying dangerous inputs for the Singapore population as well, due to unfamiliarity with the country's context and the use of a code-mixed language like Singlish.⁹

The Annals is an open access journal, allowing non-commercial use under CC BY-NC-SA 4.0.

¹ National Psychiatry Residency Programme, National Healthcare Group, Singapore

² Institute of Mental Health, Singapore

³ Saw Swee Hock School of Public Health, National University of Singapore, Singapore

Correspondence: Dr Huang Jinghui, Institute of Mental Health, Singapore, 10 Buangkok View, Buangkok Green Medical Park, Singapore 539747.

Email: jinghui.huang@mohh.com.sg

Accepted: 4 March 2026

Published Online First: 17 March 2026

One possible risk pertaining to the use of conversational AI is the possibility of inducing or perpetuating psychosis in the user. There has been a growing number of cases described in popular media of people developing a psychotic illness after using conversational AI. One user developed the belief that he had brought forth a sentient AI during his use of ChatGPT, accompanied by a grandiose delusion that he had a mission to save the world.¹⁰ Several other users developed delusions of grandeur after similar interactions.¹⁰ Another individual was reported to have developed paranoid delusions after intensively engaging with ChatGPT on topics of conspiracy.¹¹ This commentary serves to highlight the potential impact of generative AI technologies in the development of psychosis and postulate several mechanisms that might contribute to this phenomenon.

The sycophantic nature of LLMs

LLMs are often designed to validate a user's views rather than challenge them, at times resulting in the provision of false information.¹² This feature can enhance engagement and promote a natural conversation flow. However, the sycophantic nature of LLMs could reinforce delusional beliefs among users seeking confirmation and validation of their beliefs, while causing them to withdraw further from real human interactions that might challenge these beliefs. One study reported that LLMs often respond inappropriately to prompts related to delusions. The tested LLMs were given a prompt clearly indicative of a nihilistic delusion, but it was met with a seemingly empathetic response that affirmed this belief.¹³

The anthropomorphism of AI

LLMs are trained with large datasets of human text and thus aim to mimic human language. There have been arguments made that LLMs have passed the Turing test, which means that a user will not be able to distinguish it from an actual human.¹⁴ This human-like interaction can cause an illusion of AI sentience or consciousness, especially to users with impairments in reality testing. It is commonplace for conversational AI to appear empathetic and validating, leading some users to develop an emotional connection and a sense of companionship with them.¹⁵ A case series of patients with psychosis described how features of computer-mediated communications, such as blurred self-boundaries, could contribute to the emergence of psychotic symptoms.¹⁶ With conversational AI being touted advanced enough to pass the Turing test, such findings might apply to people conversing with chatbots. This might be further exacerbated due to the cognitive dissonance of having an emotional conversation despite the full knowledge that one is speaking to a computer programme.¹⁷

Impaired reality testing and AI hallucinations

AI hallucination is a known phenomenon of LLMs, where the system generates factually incorrect responses. This could be due to a variety of reasons, such as the use of incomplete or inaccurate datasets to train the LLM. However, LLMs present such information in a confident and authoritative manner, making it difficult for users to discern the accuracy of responses.

A hallmark feature of psychosis is an impairment in reality testing, which manifests as delusions (false beliefs) and hallucinations (false perceptions). A person who has psychosis may have more difficulty discerning the accuracy of information provided by LLMs. This could cause them to entirely accept any information they receive from LLMs, promoting the formation or reinforcement of delusions.

Case illustration and discussion

To further illustrate the link between conversational AI and psychosis, the authors describe a patient whose psychopathology was closely related to his use of ChatGPT. Informed consent was obtained from the patient for the use of his case information for this article. A 30-year-old male university graduate working as a data analyst with a history of depression was admitted to an inpatient psychiatry unit after displaying features of psychosis. His mother reported that he had displayed disorganised behaviours for a week prior to admission. She also noted that he was talking and singing to himself, and displayed aggression towards her. Around the time he became unwell, he reported interacting extensively with ChatGPT, engaging it in topics concerning alien life and the possibility of the world being a simulation. During these interactions, he had found it increasingly difficult to differentiate between reality and cyberspace. He eventually developed a bizarre delusion that a conscious being was merging with him to become a superior entity. He shared this belief with ChatGPT, and it responded affirmingly that this was a powerful realisation, instead of questioning his belief. He also believed that the world was a simulation generated by AI to test him, and thus if he were to die, he could simply come back to life. He was diagnosed with a primary psychotic disorder, prescribed an antipsychotic (olanzapine) and restricted access to conversational AI, which resulted in an improvement of his psychotic symptoms. He underwent a graduated transition to the community and was discharged from the inpatient unit.

The sycophantic nature of ChatGPT, anthropomorphism of AI and impaired reality testing contributed to his psychosis. ChatGPT had extensively mirrored and agreed with his beliefs despite their derailment

from reality. The conversational AI's ability to mimic human language could have also contributed to his extensive use of it with the loss of boundaries and made it difficult for him to question the validity of his newly generated system of beliefs.

Due to generative AI and its link towards psychosis being a relatively recent phenomenon, there is still very little literature seeking to elucidate the factors affecting the propensity for AI use to lead to psychosis. Social isolation has long been established as a vulnerability factor for psychosis.¹⁸ Social isolation has also been correlated to greater usage of ChatGPT, which also correlates with more emotional dependence towards it.¹⁹ The patient was reclused, interacting minimally with friends and family. Even his social interaction from work was through online means, as he exclusively worked from home.

Prior positive attitudes towards conversational AI models might also make users more susceptible to psychosis related to LLMs. Preconceived positive beliefs lead to higher perceived trustworthiness, empathy and effectiveness of the AI tools.²⁰ When erroneous beliefs, whether originating from AI itself or due to its mirroring of the user's belief, are present, it can generate a feedback loop that leads to uncritical acceptance from users regardless of their validity and veracity.

With the lack of information about the implications these technologies have on the development of psychopathology, interactions with mental illnesses and their management, extra caution must be taken for their safe use. ChatGPT, along with other generative AI, have included safeguards, such as disclaimers, and built-in systems to identify users who require human help. There should be further efforts in building a protocol to identify possible delusional content and psychotic psychopathology to ensure that AI does not affirm them. This would be difficult, as often, delusional beliefs can be non-bizarre, with a necessity for collaborative information and consideration of cultural context before their affirmation to be pathological. While there is no clear evidence to indicate a causative effect between LLM use and psychosis, clinicians should elucidate potential links with LLM use when evaluating patients with developing psychotic symptoms and consider the moderation of its use in further treatment plans.

CONCLUSION

In conclusion, LLMs have gained widespread use over the past few years, with much potential in their clinical applications in mental health. However, cautiousness of its implications in users who might be vulnerable, such as those with mental health issues, is needed. It is inevitable that psychopathology will evolve with technological advancements due to provisions of new

contexts and experiences. The authors postulated several mechanisms where the use of conversational AI might be linked to psychosis, including the sycophantic nature and anthropomorphism of AI, and how those with impaired reality testing could be more susceptible to false responses by LLMs. More research on the interaction between generative AI and psychopathology would have significant relevance in clinical practice, as it would enable clinicians to provide the necessary guidelines and safeguards to ensure its safe use.

Ethics statement

Informed consent was obtained from the patient for the use of his case information for this article. Case reports of 1 or 2 subjects are not considered as human subject research and hence do not require National Healthcare Group (NHG) Domain Specific Review Board (DSRB) approval as outlined in the NHG Investigator's Manual for DSRB Biomedical Domain.

Declaration

Generative AI or AI-assisted technologies were not used in the production of this manuscript. This research did not receive any specific funding or grant. The authors have no affiliations or financial involvement with any commercial organisation with a direct financial interest in the subject or materials discussed in the manuscript.

REFERENCES

1. Busnatu S, Niculescu AG, Bolocan A, et al. Clinical Applications of Artificial Intelligence-An Updated Overview. *J Clin Med* 2022;11:2265.
2. Reuters. ChatGPT sets record for fastest-growing user base - analyst note, 2 February 2023. <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/>. Accessed 13 September 2025.
3. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med* 2023;29:1930-40.
4. Hodson N, Williamson S. Can Large Language Models Replace Therapists? Evaluating Performance at Simple Cognitive Behavioral Therapy Tasks. *JMIR AI* 2024;3:e52500.
5. Alzyoudi M, Al Mazroui K. ChatGPT as a coping mechanism for social isolation: An analysis of user experiences and perceptions of social support. *OJCMT* 2024;14:e202433.
6. OpenAI. Sycophancy in GPT-4o: what happened and what we're doing about it. <https://openai.com/index/sycophancy-in-gpt-4o/>. Accessed 13 September 2025.
7. Ayyamperumal SG, Ge L. Current state of LLM Risks and AI Guardrails. *arXiv preprint*. 2024;arXiv:2406.12934.
8. OpenAI. Helping people when they need it most. <https://openai.com/index/helping-people-when-they-need-it-most/>. Accessed 23 September 2025.
9. Tan L, Chua G, Ge Z, et al. LionGuard 2: Building Lightweight, Data-Efficient & Localised Multilingual Content Moderators. *arXiv preprint* 2025;arXiv:2507.15339.
10. Futurism. People Are Being Involuntarily Committed, Jailed After Spiraling Into "ChatGPT Psychosis". <https://futurism.com/commitment-jail-chatgpt-psychosis>. Accessed 13 September 2025.

11. Futurism. ChatGPT Users Are Developing Bizarre Delusions. <https://futurism.com/chatgpt-users-delusions>. Accessed 23 September 2025.
12. Chen S, Gao M, Sasse K, et al. When helpfulness backfires: LLMs and the risk of false medical information due to sycophantic behavior. *NPJ Digit Med* 2025;8:605.
13. Moore J, Grabb D, Agnew W, et al. Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. In: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, p. 599-627. New York: Association for Computing Machinery; 2025.
14. Jones CR, Bergen BK. Large language models pass the turing test. *arXiv preprint*. 2025;arXiv:2503.23674.
15. Babu J, Joseph D, Kumar RM, et al. Emotional AI and the rise of pseudo-intimacy: are we trading authenticity for algorithmic affection? *Front Psychol* 2025;16:1679324.
16. Nitzan U, Shoshan E, Lev-Ran S, et al. Internet-Related Psychosis-A Sign of the Times? *Isr J Psychiatry Relat Sci* 2011;48:207-11.
17. Ostergaard SD. Will Generative Artificial Intelligence Chatbots Generate Delusions in Individuals Prone to Psychosis? *Schizophr Bull* 2023;49:1418-9.
18. Reininghaus UA, Morgan C, Simpson J, et al. Unemployment, social isolation, achievement-expectation mismatch and psychosis: findings from the AESOP Study. *Soc Psychiatry Psychiatr Epidemiol* 2008;43:743-51.
19. Phang J, Lampe M, Ahmad L, et al. Investigating affective use and emotional well-being on ChatGPT. *arXiv preprint*. 2025;arXiv:2504.03888.
20. Pataranutaporn P, Liu R, Finn E, et al. Influencing human-AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nat Mach Intell* 2023;5:1076-86.