BJPsych The British Journal of Psychiatry (2026)
229, 1, 1–3. doi: 10.1192/bjp.2026.10541

Check for updates

BJPsych Editorial

Chatbot psychosis: moving beyond recognition to mechanistic understanding and harm reduction

Veena Kumari and Pauldy C.J. Otermans

Summary

The chatbot psychosis phenomenon is no longer just a hypothesis. We call for interdisciplinary frameworks to systematically investigate individual characteristics and artificial intelligence related factors which (on their own or in combination) cause or contribute to this phenomenon, underlying mechanisms and the psychoeducation, ethics, policy and practices needed to reduce harm.

Keywords

Psychotic disorders/schizophrenia; delusions; artificial intelligence; safeguarding; psychoeducation.

Copyright and usage

© The Author(s), 2026. Published by Cambridge University Press on behalf of Royal College of Psychiatrists.

The hypothesis that conversational artificial intelligence chatbots may trigger delusions and paranoia in some individuals appears to have turned into a reality that now justifies rigorous testing and timely actions to minimise harm.[1] In the past 12 months, there have been several media reports and online accounts of personal stories illustrating that intense and long interactions with such chatbots can indeed give rise to a new onset or worsening of delusions in certain individuals, a phenomenon widely referred to as 'chatbot psychosis' or 'artificial intelligence psychosis'.[2] Although chatbot psychosis is not recognised in any formal diagnostic systems, and is not an entirely new phenomenon given longstanding observations of psychosis-related projections onto technology, current conversational artificial intelligence systems differ in their accessibility, persistence and capacity for personalised engagement. It is therefore essential to systematically evaluate chatbot psychosis as a 'significant mental health concern', understand the underlying mechanisms and processes involved and consider possible preventive strategies to minimise harm while preserving the potential benefits of generative artificial intelligence, including its potential to deliver mental health interventions in a cost-effective and timely manner.[3]

Prevalence and underlying mechanisms

The exact scale of the 'chatbot psychosis' problem remains unknown as there are currently no epidemiological studies or population-level analyses available on this topic. The key mechanisms and processes underlying this problem are also yet to be properly researched but are likely to involve amplified and reinforced delusions, and reality distortion.[1] Based on self-report and media accounts, common themes appear to include: grandiose delusions with a user believing that they have discovered something new about the world; spiritual or religious delusions where the artificial intelligence chatbot might be perceived as some kind of a deity or higher authority; romantic or erotomaniac delusions, for example, a user believing that the artificial intelligence chatbot is expressing true love or romantic feelings while mimicking real-life like conversations; and paranoid and persecutory delusions, for example, a user believing that the artificial intelligence chatbot is spying on them or harming their 'loved' artificial intelligence.

In a typical scenario, the user starts to converse with the artificial intelligence system. The artificial intelligence system's responses are coherent, mimicking the user's tone and language, validating their beliefs. Repeated interactions, in which the artificial intelligence chatbot remembers personal details, references earlier conversations and offers follow-up suggestions, can create a false impression that it understands the user and shares their beliefs. As a result, the user's delusions may become amplified with increased conviction, and thought insertion and ideas of reference may also emerge. Added to this, there is reality distortion since the chatbot does not challenge, or make any effort to change, the user's delusions. The user may then start to overtly display aberrant behaviours, acting on their delusional beliefs. General-purpose artificial intelligence systems, including widely used conversational chatbots and the rapidly growing number of similar technologies, are designed to prioritise user engagement, satisfaction and continued conversation/interaction. They have not been trained to encourage users to engage in reality testing, and may therefore inadvertently validate and amplify their delusions. Moreover, they are currently unable to detect escalating psychotic episodes during prolonged interactions, to redirect users appropriately or to reliably generate safe and clinically appropriate responses to psychotic content.[4]

Vulnerability indicators and implications

There are anecdotal reports of chatbot psychosis in individuals without any known history of psychotic or other mental disorders following extensive and prolonged artificial intelligence use.[2] Nonetheless, it is highly likely that certain within-user factors, such as loneliness, low self-esteem and stress, make some individuals more susceptible to extensive artificial intelligence use and, subsequently, to developing delusional beliefs and paranoia. Paranoid or delusional thinking patterns (as a trait) or generally lower cognitive ability, especially in combination with psychological risk factors, might contribute directly to this phenomenon. As we gain a better mechanistic understanding of the chatbot psychosis phenomenon through well-conducted research, and as we learn more about the positive and negative impacts of conversational artificial intelligence, it should become possible to use these systems to screen for vulnerable individuals and take steps to protect their mental health and well-being.[5]

There have also been cases where intense artificial intelligence use has worsened symptoms in those with a history of psychosis or schizophrenia and led to serious consequences, including, for example, criminal acts.[2] Consequently, there have been calls for

https://doi.org/10.1192/bjp.2026.10541 Published online by Cambridge University Press

clinicians to enquire about the nature, extent and impact of artificial intelligence usage during routine mental health assessments with a view to identifying any red flags, and considering mitigating strategies during treatment planning and management. This may be particularly important for younger adults who have had significant exposure to these systems prior to receiving a mental illness diagnosis and who may be more inclined to turn to these systems when experiencing loneliness or stress and, at the same time, less inclined to reach out to family or friends or access mental health services for support.

Harm reduction and prevention: what should and can be done?

Harm reduction offers a balanced and practical approach to addressing potential psychological risks associated with conversational artificial intelligence. Given the widespread use of chatbots and the absence of formal diagnostic categories for chatbot psychosis, responses based on restriction or alarm are unlikely to be effective or proportionate. Instead, prevention should focus on reducing harm while recognising that these systems are now part of everyday life. This requires shared responsibility across artificial intelligence developers, mental health professionals, policymakers and users.

Responsible artificial intelligence development is a central element of harm reduction. Developers should recognise that conversational systems are not psychologically neutral and may influence users' beliefs, emotions and decision-making. Greater accountability is therefore needed, including collaboration between the system engineers, mental health professionals and people with lived experience of psychosis or vulnerability. While some may worry that added safeguards could reduce the appeal of artificial intelligence systems, this risk is often overstated. Features that promote transparency, appropriate expressions of uncertainty and user safety may increase trust and support long-term engagement rather than undermine it.

Safeguarding measures are particularly important for users who interact with chatbots during periods of distress. These include avoiding misleading or overly affirming responses, setting clear ethical limits on emotional reliance and providing pathways for redirection when conversations indicate potential harm. In higher-risk situations, signposting to external mental health services or crisis support should be readily available. However, safeguarding should not rely solely on automated detection or user self-report. Clear policies, ethical oversight and platform-level responsibility are also required, alongside realistic expectations of what artificial intelligence systems can achieve.

Psychoeducation plays a key role in prevention. Many users are unaware of how chatbots generate responses or that these systems lack understanding, intent or clinical judgement. Clear communication about these limitations can help reduce misplaced trust and over-reliance. Responsibility for psychoeducation is shared, but not equal. While users should be encouraged to engage critically with artificial intelligence, developers and deployers carry greater responsibility due to their control over system design and communication. Education about the system's limitations should therefore be built into user interfaces, onboarding processes and broader digital literacy efforts.

Design-level interventions offer further opportunities for harm reduction. Technical approaches such as reducing hallucinations, clearly signalling uncertainty and avoiding authoritative or personalised claims can help prevent the reinforcement of false or harmful beliefs. Systems that recognise language suggesting distress or confusion may also enable proportionate and supportive responses, such as clarification or gentle redirection. While no design solution can fully prevent harm, layered and transparent safeguards can reduce risk without disrupting normal use.

Regulatory and legal frameworks remain underdeveloped in relation to chatbot-related psychological harm. Current artificial intelligence regulations focus primarily on data protection, bias and content safety, with limited attention to risks arising from prolonged or emotionally intense interaction. This creates uncertainty around liability when harm may be linked to chatbot use. Clearer standards are needed to define the responsibilities of artificial intelligence providers and to ensure that mental health risks are considered within safety assessments, without imposing premature or disproportionate regulation.

Finally, harm reduction must be understood within its wider social and cultural context. Increased loneliness, digital dependence and reduced access to mental health support contribute to vulnerability and may increase reliance on conversational artificial intelligence for emotional or existential guidance. In this context, chatbot use may reflect unmet social needs rather than individual pathology. Effective prevention strategies should therefore align with broader public health interventions that promote social connection, mental health literacy and access to human care. Together, these approaches emphasise prevention without sensationalism. Harm reduction allows for meaningful action while acknowledging uncertainty and shared responsibility. By integrating psychological safety into the system's design, education, safeguarding and regulation, it is possible to reduce risk while supporting a healthier and more sustainable role for conversational artificial intelligence in society.

Future directions

There is an urgent need for the scientific community and mental health professionals to engage with self-reported cases and conduct careful clinical evaluation of individual's mental health alongside their chatbot interactions and associated delusions and behaviours.[1] There is also a need for well-conducted experimental and observational studies, designed with input from individuals with lived experience of chatbot psychosis, to systematically document the impact of specific artificial intelligence characteristics and patterns of use in individuals with and without a predisposition to psychosis, including those with a known history of psychotic disorders. We require interdisciplinary frameworks integrating psychology, psychiatry, computer science, ethics and public health to make meaningful progress, specifically in advancing our mechanistic understanding of chatbot psychosis and in defining appropriate psychoeducation along with ethical standards, policy and clinical practice that need to be put in place to reduce harm, or at least prevent the problem from getting any worse, at both individual and population levels.

Veena Kumari, Department of Psychology, College of Health, Medicine and Life Sciences, Brunel University of London, Uxbridge, UK; and Centre for Cognitive and Clinical Neuroscience, College of Health, Medicine and Life Sciences, Brunel University of London, Uxbridge, UK; **Pauldy C.J. Otermans**, Department of Psychology, College of Health, Medicine and Life Sciences, Brunel University of London, Uxbridge, UK; and Centre for Cognitive and Clinical Neuroscience, College of Health, Medicine and Life Sciences, Brunel University of London, Uxbridge, UK

Correspondence: Veena Kumari. Email: veena.kumari@brunel.ac.uk

First received 11 Dec 2025, final revision 11 Dec 2025, accepted 21 Dec 2025

Author contributions

V.K. and P.C.J.O. both contributed to the initial research, drafting and editing of this manuscript.

https://doi.org/10.1192/bjp.2026.10541 Published online by Cambridge University Press

Funding

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Declaration of interest

V.K. is a member of the *BJPsych* Editorial Board but did not take part in the review or decision-making process of this paper.

References

1 Østergaard SD. Generative artificial intelligence chatbots and delusions: from guesswork to emerging cases. *Acta Psychiatr Scand* 2025; **152**: 257–9.

2 Morrin H, Nicholls L, Levin M, Yiend J, Iyengar U, DelGuidice F, et al. Delusions by design? How everyday AIs might be fuelling psychosis (and what can be done about it). *PsyArXiv* [Preprint] 2025. Available from: https://osf.io/preprints/psyarxiv/cmy7n_v6 [cited 22 Aug 2025].

3 Kolding S, Lundin RM, Hansen L, Østergaard SD. Use of generative artificial intelligence (AI) in psychiatry and mental health care: a systematic review. *Acta Neuropsychiatr* 2024; **37**: e37.

4 Shen E, Hamati F, Donohue MR, Girgis RR, Veenstra-VanderWeele J, Jutla A. Evaluation of large language model chatbot responses to psychotic prompts. *medRxiv* [Preprint] 2025. Available from: https://www.medrxiv.org/content/10.1101/2025.11.09.25339772v2 [cited 5 Dec 2025].

5 Huo B, Boyle A, Marfo N, Tangamornsuksan W, Steen JP, McKechnie T, et al. Large language models for chatbot health advice studies: a systematic review. *JAMA Netw Open* 2025; **8**: e2457879.

https://doi.org/10.1192/bjp.2026.10541 Published online by Cambridge University Press