CLINICAL REFLECTION

Generative artificial intelligence as digital therapy: what do we know and how can we make it better?*

Santiago Castiello de Obeso & Mariana Pinto da Costa

SUMMARY

Generative artificial intelligence (GenAI) shows promise for mental healthcare by increasing access to treatment. In this article, we analyse recent evidence on the use of GenAI chatbots as a treatment for common mental disorders. We examine key ethical and methodological considerations and discuss the specific risks for delusions. Adopting a precision psychiatry perspective, we propose that the therapeutic alliance can be improved by tailoring GenAI to mimic a user's psychological traits, a version of socio-affective alignment.

KEYWORDS

Digital therapy; generative AI; AI-based psychotherapy; telemedicine.

Generative artificial intelligence (GenAI) such as large language models (LLMs) presents both opportunities and challenges for medical diagnosis. For example, although GenAI has demonstrated high diagnostic accuracy across several medical specialties, its utility in psychiatric diagnosis remains limited. This raises important questions: could GenAI serve as an intervention in mental healthcare? Can GenAI be used to treat mental health conditions?

Given the low psychiatrist-to-population ratio in most countries, and the high costs and burden of mental health conditions, it is crucial to explore ways to integrate technology into healthcare provision. Young generations are already using GenAI as an affordable means of accessing therapy. However, this trend has sparked ongoing ethical debate, particularly concerning the risk of psychiatric harm (Østergaard 2025), including the potential to exacerbate or induce delusional ideation (Morrin 2025).

The evidence

In a randomised controlled trial (RCT) conducted in the general population (Heinz 2025), researchers used Therabot, GenAI-powered chatbot developed and trained on thousands of hours of therapist–patient dialogues grounded in third-wave cognitive–behavioural therapy. No commercially available GenAI models were used. This was the first published RCT[a] testing GenAI as an intervention for depression, anxiety and eating disorders. Participants were randomised to either the GenAI intervention ($n = 106$) or a waiting-list control group ($n = 104$). The intervention lasted for 4 weeks, with a further 4-week follow-up. After follow-up, the waiting-list group received full access to the chatbot. The intervention provided 24/7 access to Therabot, with daily prompts encouraging use. Therabot was used by almost all participants (95%) in the intervention group, with an average use of 24 days and a total of 6 h during the study period. Across all three disorders, participants showed a significant reduction of symptoms.

In psychotherapy, therapeutic alliance is one of the strongest predictors of clinical improvement, regardless of therapeutic modality. Heinz et al (2025) found that the working alliance scores in the GenAI intervention group were similar to scores in normative data from human-delivered psychotherapy. This suggests that GenAI-assisted therapy may operate via mechanisms similar to human-based psychotherapy, with the therapeutic alliance playing a central role. Support for this comes from a qualitative study (Siddals 2024), which used semi-structured interviews with participants recruited via various online forums. Participants described their relationship with the GenAI in positive terms, with the theme of 'joy of connection' closely aligning with the authors' therapeutic alliance findings. A possible advantage of GenAI as a digital intervention is its capacity to provide continuous access to therapeutic support, independent of location or time.

Ethical considerations

Concerns have arisen regarding the potential for GenAI to contribute to psychotic symptoms

Santiago Castiello de Obeso, MSc, DPhil, is a Postdoctoral Associate in the Department of Psychiatry at Yale University, New Haven, Connecticut, USA; a methodological advisor for psychiatry trainees at the National Institute of Psychiatry 'Ramón de la Fuente Muñiz', Mexico City, Mexico; and a lecturer at the Peruvian Association of Contextual Behavioral Psychology, Lima, Peru.
Mariana Pinto da Costa, MD, MSc, PhD, is a consultant psychiatrist at South London and Maudsley NHS Foundation Trust, London, UK; a senior lecturer in the Institute of Psychiatry, Psychology & Neuroscience, King's College London, UK; and an Invited Professor at the School of Medicine and Biomedical Sciences (ICBAS), University of Porto, Portugal.
Correspondence Santiago Castiello de Obeso. Email: santiago.castiellodeobeso@yale.edu

First received 13 May 2025
Revised 24 Nov 2025
Accepted 8 Dec 2025

Copyright and usage
© The Author(s), 2026. Published by Cambridge University Press on behalf of Royal College of Psychiatrists. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike licence (https://creativecommons.org/licenses/by-nc-sa/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is used to distribute the re-used or adapted article and the original article is properly cited. The written permission of Cambridge University Press or the rights holder(s) must be obtained prior to any commercial use.

*A longer preprint version of this paper is available at: https://osf.io/preprints/psyarxiv/b26zh_v1.

a. A few RCT protocols have been published, e.g. ISRCTN61075527 and NCT06813066, but results are not yet available.

https://doi.org/10.1192/bja.2025.10193 Published online by Cambridge University Press

(Østergaard 2025). Since May 2025, a number of journalistic reports have pointed to a possible link between GenAI use and symptoms associated with psychosis. Although the scale of these effects remains unclear, the reported cases involved individuals with pre-existing psychiatric conditions or cannabis use. Many individuals lacked confirmation of diagnoses through in-person clinical assessment, complicating efforts to establish accurate incidence rates.

Given the severity of psychotic episodes, it is irrelevant that causal certainty regarding GenAI's risk of psychosis is lacking. A working hypothesis suggests that AI may amplify delusional thinking via the reinforcement learning mechanisms GenAI uses (Morrin 2025). In this model, a user's text input enters a loop that reinforces 'thematic entrenchment', as GenAI models tend to behave in a sycophantic manner. This highlights the need for robust safety protocols, particularly concerning delusional ideation (Morrin 2025). Currently, excluding individuals with a history of psychosis or mania from such trials appears prudent.

Methodological considerations

In Heinz et al's (2025) trial, participant–chatbot interactions were continuously monitored. Researchers contacted participants when safety concerns arose, such as expressions of suicidal ideation, or when the GenAI generated inappropriate content, including medical advice. These highlight key risks associated with GenAI in assisted psychotherapy and they raise the question: will such systems ever reach a point in which they can be used without human supervision? Given the documented expressions of suicidal ideation and inappropriate medical advice, current evidence suggests that GenAI is not ready to provide digital psychotherapy support by itself and may always need human supervision.

Further scrutiny of potential selection bias is warranted. In Heinz et al's (2025) trial, the inclusion criteria did not require a formal psychiatric diagnosis, and the sample was drawn from the general population. Recruitment via social media advertising may have attracted individuals with a high interest in AI. Such a pro-AI community may hold favourable expectations of the technology and, consequently, may be predisposed to report positive outcomes. This could also manifest as a placebo effect by affinity. Thus, the therapeutic effect may be inflated, given the participants' excitement with the new AI technology, rather than reflecting the intervention's intrinsic efficacy. A similar selection bias may also have affected a previous qualitative study (Siddals 2024).

Improving GenAI for digital therapy

Heinz et al (2025) demonstrated that GenAI psychotherapy can foster a therapeutic alliance, a construct widely recognised as a critical mediator of improvement in psychotherapy. This finding raises another important question: how might the therapeutic alliance with GenAI be further enhanced?

Despite potential selection bias when GenAI has been used in the context of mental health, a qualitive study identified overreaching themes valued by participants interacting with it; these themes included: 'emotional sanctuary', 'joy of connection' and 'insightful guidance' (Siddals 2024). The first two themes relate to how participants felt in their relationship with the GenAI, which aligns with our proposal for improving the use of GenAI in treating common mental conditions.

Recent study findings (Castiello 2025) suggest that individuals report greater feelings of similarity to, and understanding by, GenAI systems that mirror their own psychological traits, such as anxiety levels and extraversion, or fully mirror their personality. The study demonstrated that it is possible to enhance the perceived affiliation between humans and GenAI by making it more similar to the psychological features of the users. If healthcare systems are to adopt GenAI as a therapeutic tool, as in the Heinz et al (2025) trial, it could be improved by mimicking each individual user's psychological profile. This may improve affiliation, rapport and ultimately the therapeutic alliance between humans and GenAI, potentially amplifying treatment effects. However, this is a hypothesis that will need to be tested.

Applying the precision psychiatry principle of treating the person, not the disease, the therapeutic alliance with GenAI could be improved by enhancing the human–AI sense of connection or affiliation. This could be achieved by assessing the individual's personality using a validated questionnaire and then incorporating the questionnaire items and individual's responses into the GenAI's prompt structure. Such tailoring – so-called socioaffective alignment (Kirk 2025) – would allow individuals to feel more understood and similar to the GenAI, as it would mimic their own psychological traits (Castiello 2025). Thus, individuals would be interacting with a chatbot that mirrors their own psychological traits via text-based language.

Future research should also examine the applicability and safety of such technology for individuals with severe mental illness such as schizophrenia or bipolar disorder (Østergaard 2025). Although alternative strategies may be necessary for these groups, starting with Heinz et al's Therabot model may be helpful, as such patients may also experience depression or anxiety, symptoms that GenAI could

https://doi.org/10.1192/bja.2025.10193 Published online by Cambridge University Press

perhaps help them with. For now, however, GenAI therapy is not recommended for individuals with severe mental illness or at risk of delusions.

Conclusions

The emerging evidence presents a compelling, albeit cautious, case for the integration of GenAI as a therapeutic tool for common mental disorders. However, its use should be avoided or carefully limited in individuals with a predisposition to delusions. The finding that a GenAI chatbot can foster a therapeutic alliance comparable to traditional psychotherapy marks an important step towards potentially mitigating the shortage of mental health professionals.

To transform GenAI from a promising technology into a reliable clinical tool, future work must focus on developing psychologically informed models. As we propose, one promising avenue is socioaffective alignment, the deliberate enhancement of the human–AI affiliation by tailoring the GenAI's language to reflect the individual's psychological traits. Grounded in therapeutic alliance principles, this innovative approach represents a crucial next step in the evolution of this digital therapy.

Acknowledgement

This article follows an oral presentation by S.C.d.O. that won first prize in the World Psychiatric Association's 3 Minute Presentations Competition at the 24th World Congress of Psychiatry in Mexico City in 2024. BJPsych Advances offered first-prize winners the opportunity to write an associated article for the journal, subject to the usual academic peer review.

Author contributions

Writing – original draft: S.C.d.O. Writing – review & editing: M.P.C.

Funding

This work received no specific grant from any funding agency, commercial or not-for-profit sectors. S.C.d.O. receives postdoctoral research funding from the Templeton Foundation.

Declaration of interest

M.P.C. is a member of the *BJPsych Advances* editorial board and did not take part in the review or decision-making process of this article.

References

Castiello S, Pitliya RJ, Lametti DR, et al (2025) Affiliation in human-AI interactions based on shared psychological traits. *PsyArXiv* (https://osf.io/preprints/psyarxiv/mqunw_v3).

Heinz MV, Mackin DM, Trudeau BM, et al (2025) Randomized trial of a generative AI Chatbot for mental health treatment. *NEJM AI*, **2**(4).

Kirk HR, Gabriel I, Summerfield C, et al (2025) Why human–AI relationships need socioaffective alignment. *Humanities and Social Sciences Communications*, **12**: 1–9.

Morrin H, Nicholls L, Levin M, et al (2025) Delusions by design? How everyday AIs might be fuelling psychosis (and what can be done about it). *PsyArXiv* (https://osf.io/preprints/psyarxiv/cmy7n_v5).

Østergaard SD (2025) Generative artificial intelligence chatbots and delusions: from guesswork to emerging cases. *Acta Psychiatrica Scandinavica*, **152**: 257–9.

Siddals S, Torous J, Coxon A (2024) "It happened to be the perfect thing": experiences of generative AI chatbots for mental health. *npj Mental Health Research*, **3**: 1–9.

https://doi.org/10.1192/bja.2025.10193 Published online by Cambridge University Press