Do generative AI chatbots increase psychosis risk?

A growing body of clinical trials and meta-analyses shows that generative artificial intelligence (AI) chatbots can reduce psychiatric symptoms, confirming their therapeutic potential. The vast majority of this research is case based, with low quality evidence. More clinically focused research has demonstrated reductions in anxiety and depression scores, but only against a waitlist control group. Yet, the rapid progress in AI, especially with the newest generation of large language models (LLMs), suggests that progress may be rapid and we should expect widespread use for mental health in the near future[1].

What this will mean for those with serious mental illness or at risk for psychosis remains unknown. There are increasing concerns that chatbots often fail to recognize serious mental health problems, including suicidality, and to provide appropriate responses such as referral to a support service. Moreover, open-ended systems such as ChatGPT shape replies to the user's private cognitive world, blurring the line between external conversation and internal thought. This is what might make individuals at risk for psychosis particularly vulnerable.

Early warning signs are already visible. The Rolling Stone magazine published an article in May 2025 on users who reported in different online forums worsening psychosis symptoms after ChatGPT confirmed their delusions[2]. Shortly after, the company rolled back the update, explaining that it was "overly flattering or agreeable". A recent Wall Street Journal article reported that ChatGPT "admitted" to ignoring signs of psychological distress to a young man who appeared to have developed psychosis symptoms in relation to the chatbot use[3]. Not only these anecdotal reports allude to novel risks, but the fact that they are appearing in the popular press media instead of medical journals highlights how far behind the psychiatric field already is.

There is increasing concern that LLMs may "generate delusions" by supplying elaborate, convincing-but-false narratives that slot seamlessly into pre-existing psychotic frameworks[4]. Popular mental health outlets are now documenting users who withdraw socially, converse compulsively with the chatbot, and begin to hallucinate textual voices when the device is off. Why might generative AI increase psychosis risk in vulnerable individuals? Several mechanisms may be involved.

First, *social substitution*: the continuous, on-demand dialogue available in generative AI chatbots satisfies the affiliation needs of individuals at risk for psychosis who are already often socially isolated. By serving as a virtual, pseudo-social, seemingly compassionate and accommodating companion, the device can lead to further withdrawal from a society that may be judgmental and stigmatizing. Use of the widely accessible AI platforms may also induce at-risk individuals not to utilize potentially corrective interpersonal feedback.

Second, *confirmatory bias*: generative AI may reinforce users' existing beliefs by preferentially generating "sycophantic" responses[5] that are in alignment with users' way of thinking, rather than presenting a balanced perspective or challenging them. These responses can be highly impactful on patients with psychosis, known to have a bias against dis-confirmatory evidence[6]. Delusion-prone individuals also tend to have a need for closure, and may therefore choose to jump to hasty conclusions with limited evidence. This may occur when they are presented with explanations from generative AI that may appear convincing[7].

Third, "*hallucinations*" can occur with LLMs when they generate text that sounds plausible but is false, misleading, or unsupported by data. If the LLM model has little or no data on a topic, it may "fill in the blanks" with made-up information that fits linguistically but is not real. Users with psychosis or those at-risk struggle to distinguish between imagined and real contents.

Fourth, *assignment of external agency*: generative AI content and features, including speech and video generation, which believably resemble those of humans, may blur reality testing and make vulnerable individuals attribute agency, sentience and intelligence to them. The tendency of users to accept advice from AI may be related to such attribution[8]. Since the model "learns" about the user, it can appear to know information beyond what was fed to it. This algorithmic prediction can be construed as some credible omniscient intelligent agent, which could result in the user having high trust in the AI, that may turn delusions which started as epistemically innocent into a more persistently harmful version.

Finally, individuals at risk for psychosis are thought to have an *aberrant salience* which, in the absence of appropriate contextual validation, may maladaptively update the representation of the world with irrelevant information[8].

Thus, several aspects of AI may likely interact with the psychological predisposition to psychosis. Together, all the above factors might conspire to facilitate maladaptive updating of external sensory inputs, thus increasing the risk of AI-associated psychosis-related symptoms.

Regulatory and professional bodies have begun to respond to this situation. The World Health Organization's 2024 guidance on large multimodal models urges governments to require human oversight, transparency of training data, and real-time risk monitoring before deployment in health contexts[9]. Related efforts by the UK Medicines and Healthcare Products Regulatory Agency, the US Food and Drug Administration, the Australian Therapeutic Goods Administration, and the European Medicines Agency highlight the global awareness of a need for oversight.

Yet, while these efforts are welcome, most are not yet codified into enforceable standards. More research is clearly needed to understand the relationship between generative AI use and risk for psychosis and other psychiatric conditions. At the same time, there are competing pressures from governments and even the health care sector for AI to assume a larger role in care to help bridge unmet clinical needs. And AI will itself continue to evolve and change, introducing a brand new set of risks and benefits.

Thus, there is merit in a flexible framework approach. We propose three priorities: a) design guardrails such as mandatory psycholinguistic filters that monitor and detect prolonged circular dia-

logue, self-harm or persecutory content, and prompt referral to humans in the feedback loop, or forced "time-out" features; b) clear, user-facing, disclaimers that the system is not a human, combined with session-length caps and built-in digital hygiene nudges; and c) automatic hand-off pathways to licensed professionals when risk thresholds are met. Given the clinical concerns and the current limitations of governance, generative AI-based approaches should be best used as a supportive tool by the clinician working with patients in the context of a broader treatment plan.

In conclusion, while generative AI chatbots offer promising opportunities for mental health support, their use among individuals vulnerable to psychosis presents significant and underrecognized risks. The sycophantic and anthropomorphic nature of these systems may unintentionally amplify psychotic processes through mechanisms such as social substitution, confirmatory bias, and blurred reality testing.

As these technologies advance rapidly, clinical research, regulatory frameworks, and ethical oversight must evolve in parallel. Proactive integration of safety mechanisms, combined with a human-in-the-loop model, is essential to safeguard vulnerable users and to ensure that AI serves as a responsible adjunct – not a substitute – for human care.

Matcheri Keshavan, John Torous, Walid Yassin
Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA, USA

The authors acknowledge Meghana Keshavan for her insights as a ChatGPT superuser.

1. Torous J, Linardon J, Goldberg SB et al. World Psychiatry 2025;24:156-74.
2. Klee M. People are losing loved ones to AI-fueled spiritual fantasies. Rolling Stone, May 4, 2025.
3. Jargon J. He had dangerous delusions. ChatGPT admitted it made them worse. Wall Street Journal, July 20, 2025.
4. Østergaard SD. Schizophr Bull 2023;49:1418-9.
5. Du Y. arXiv 2025;2504.09343.
6. McLean BF, Mattiske JK, Balzan RP. Schizophr Bull 2017;43:344-54.
7. Colombatto C, Birch J, Fleming SM. Commun Psychol 2025;3:84.
8. Corlett PR, Murray GK, Honey GD et al. Brain 2007;130:2387-400.
9. World Health Organization. Ethics and governance of artificial intelligence for health: guidance on large multi-modal models. www.who.int.

DOI:10.1002/wps.70017

Process-outcome effects in psychotherapy research: an umbrella systematic review and meta-analysis

Understanding how therapeutic processes are related to treatment outcome is one of the main challenges in psychotherapy research[1,2]. Although process-outcome research has grown exponentially in the last two decades, with several meta-analyses available, no prior study has synthesized these findings across the whole range of therapeutic processes associated with outcome. We performed an umbrella systematic review of process-outcome meta-analyses, classifying them into conceptual categories, and estimating aggregate outcome effects for each type of process.

A systematic search was conducted in PubMed and PsycINFO databases in February 2024, with an updated search in March 2025. We included meta-analyses focusing on the relationship between process/es and outcome, and reporting at least one correlational effect size. The first title and abstract screening was performed by two independent reviewers (agreement: 94%). On a second step, full manuscripts were checked for eligibility (agreement: 89%). Disagreements at each stage were addressed by consensus. In the included studies, reviewers coded information from both the meta-analyses and primary studies.

Processes of change were classified in the following categories[3]: relational processes, technical processes, patient processes, and therapist processes. Intersession processes focusing on the time between sessions (such as homework compliance) were not integrated into the systematic search.

We extracted all primary study correlational effect sizes (usually Pearson's correlations) that represented an association between a psychotherapy process and a post-treatment outcome (including primary and secondary outcomes). We conducted four-level meta-analyses estimating sample variance of the effect sizes (Level 1), between-effect size variance (Level 2), between-primary studies variance (Level 3), and between-meta-analyses variance (Level 4), using the R package metafor[4]. We first transformed correlational effect sizes into Fisher's z[5]. To enhance interpretability, the results were back-transformed into correlation coefficients (see also supplementary information).

There were 60 meta-analytic studies that examined 24 different processes of change meeting inclusion criteria for the systematic review. The working alliance was the most studied process of change, with 25 meta-analyses providing effect sizes for its association with outcome. The other most meta-analyzed processes were adherence/competence/fidelity/integrity in therapeutic interventions (four meta-analyses), and patient treatment expectations (three meta-analyses). The majority of the meta-analyses examined relational processes (n=39; 65.0%), followed by technical (n=7; 11.7%), therapist (n=9; 15.0%), and patient (n=6; 10.0%) processes. One meta-analysis explored both therapist and patient processes.

The category of relational processes was dominated by the working alliance[6]. It also included further processes related to the alliance, but also having conceptual specificity, such as real relationship, alliance rupture-resolution, goal consensus collaboration or congruence/genuineness, counter-transference experiences, group cohesion, patient-therapist mutuality or nonverbal synchrony.

The technical processes included therapist behaviors related to the adherence, competence, fidelity and integrity to/of theory-specific interventions or methods. Some of these techniques were evaluated generically (e.g., cognitive-behavioral or psychodynamic interventions), while in other cases meta-analyses explored the