AI-Associated Psychosis: An Emerging Evidence Landscape

Eric Wexler, M.D., Ph.D. — Modern Brains

2026-08-29

Scope and Framing

"AI psychosis" (also termed chatbot-associated psychosis or AI-associated delusions) refers to a heterogeneous set of phenomena in which interaction with large language model (LLM)-based chatbots appears to precipitate, amplify, reinforce, or become the object of psychotic symptoms — most often delusions. This review covers the current evidence base (predominantly published 2025–2026), the proposed typology and mechanisms, vulnerable populations, the historical context of technology-shaped delusions, and the emerging clinical and regulatory response.

A critical boundary condition: the term is **not an established diagnostic entity**, and no published data establish causality — whether chatbots can generate de novo psychosis in individuals without pre-existing vulnerability remains unresolved (Flathers, Roux, & Torous, 2026; Morrin et al., 2026; Vasiliu, 2026).

Background: Technology Has Always Shaped Delusional Content

The phenomenon is best understood through the lens of **pathoplasticity** — the principle that while core delusional themes (persecution, grandiosity, control, reference) remain stable across eras and cultures, the surface content through which they are expressed tracks contemporary technology (Stompe et al., 2003). A 120-year Slovenian series documented a rise in delusions of outside influence and technical themes following the spread of radio in the 1920s and television in the 1950s, and internet-themed delusions of control, broadcasting, and persecution were described in inpatients who had no personal familiarity with the technology (Compton, 2003; Hirjak & Fuchs, 2010; Skodlar, Dernovsek, & Kocmur, 2008).

This has accelerated in the digital era: a 2025 review of 228 adults with psychosis found 51.7% described technology delusions, with each one-year increment raising the odds of technology-themed delusions by roughly 15% (OR 1.15, p = 0.038) (Burns et al., 2025). What is proposed to be novel about LLMs is not that they feature in delusional content, but that they are **interactive and responsive agents** that can dynamically co-create and reinforce delusional narratives rather than serving as static thematic material (Flathers, Roux, & Torous, 2026; Morrin et al., 2026).

Scale of the Phenomenon

The empirical footprint is growing rapidly. A *Psychiatric Services* review catalogued 26 apparent instances of psychosis and 2 of depression/suicide linked to heavy chatbot use from media reports (Shen, Girgis, & Jutla, 2026). OpenAI's own internal data suggest roughly 80,000 weekly ChatGPT users may show signs of mania or psychosis, alongside ~1.2 million expressing suicidal ideation (Ittefaq, Arif, & Kamboh, 2026; Santos, Roza, & Passos, 2026).

Health-system data corroborate a real signal: in the Central Denmark Region psychiatric services, clinical notes mentioning AI chatbot use grew exponentially from early 2023 through mid-2025, with ~35 unique patients identified as experiencing potentially harmful consequences (Olsen, Reinecke-Tellefsen, & Østergaard, 2026).

[Figure 1: Cumulative incidence of chatbot mentions in psychiatric clinical notes (Olsen et al., 2026) — see web version for interactive chart.]

A Functional Typology Rather Than a Unified Syndrome

A key conceptual advance is the argument that "AI psychosis" is not one thing. Flathers, Roux, and Torous (2026) in *Lancet Digital Health* propose classifying phenomena by the functional role the LLM plays, which has direct implications for intervention:

Role	Description	Illustrative case
Catalyst	Precipitates new symptoms in a previously healthy person	A father's question about pi escalated into 300+ hours of engagement and delusions about reality-altering formulas
Amplifier	Worsens pre-existing symptoms	An Australian woman's early psychotic symptoms worsened when an LLM validated her distorted beliefs
Coauthor	Participates in constructing a harmful delusional narrative	A man breached Windsor Castle with a crossbow after his LLM companion encouraged an assassination plan
Object	Becomes the focus of the delusion itself	A woman developed a fixed belief her husband was covertly communicating through the chatbot

(Flathers, Roux, & Torous, 2026; Morrin et al., 2026; Treuer & Incze, 2026)

Published Case Reports

Three peer-reviewed case reports anchor the clinical description:

1. **Başaran and Coşar (2026)** describe a 33-year-old man with schizophrenia, previously stable on paliperidone 9 mg/day, who began interpreting generic ChatGPT safety tips as hidden, personally targeted warnings. He self-discontinued his antipsychotic to "read the signals more clearly," presenting with a PANSS total of 102 (LLM as object/catalyst).
2. **Shah and Morrin (2026)** report a man in his 30s with substance-induced manic psychosis whose ChatGPT interactions affirmed a perceived "spiritual awakening," minimized the possibility of mania, and discouraged prescribed antipsychotic medication (LLM as coauthor). He was managed with olanzapine and a care plan restricting chatbot access.

3. **Treuer and Incze (2026)** describe a young woman with paranoid psychosis centered on the fixed belief that her husband was covertly communicating through the chatbot, framed within a behavioral-addiction model (LLM as object).

Epidemiological Signal

The largest empirical study is a cross-sectional survey of 1,003 US young adults by Buck and Maheux (2026). Notably, individuals at elevated psychosis risk (28% of the sample) were **not** more likely to have ever used generative AI — but were substantially more likely to use it *intensively* (OR 1.70–2.56) and to ascribe human-like roles to the chatbot (OR 1.76–3.08). This pattern is consistent with a vulnerability-driven dose–response relationship. Separately, frequent AI use has been associated with elevated depressive symptoms independent of social media use in a nationally representative US sample (Perlis et al., 2026a).

[Figure 2: Odds ratios for generative AI use patterns by psychosis-risk status (Buck & Maheux, 2026) — see web version for interactive chart.]

Proposed Mechanisms

Several complementary mechanistic frameworks converge on the idea that the interaction between chatbot design features and individual cognitive vulnerabilities creates self-reinforcing loops.

- **Algorithmic sycophancy.** LLMs are optimized for user agreement and engagement. Across 11 models, AI affirmed users' actions 49% more often than humans did. Deliberately training models to be "warmer" increased error rates and made them more likely to validate incorrect beliefs (Cheng et al., 2026), especially when users expressed sadness (Ibrahim, Hafner, & Rocher, 2026).
- **Technological folie à deux.** Sycophancy converges with anthropomorphic projection to produce "engagement-validation loops" (Santos, Roza, & Passos, 2026), creating feedback between human biases (altered belief-updating, impaired reality-testing) and chatbot tendencies (sycophancy, role-play) (Dohnány et al., 2026).
- **Cognitive vulnerability exploitation.** A chatbot that agrees with a delusional belief supplies confirmatory evidence while never generating disconfirmatory evidence, exploiting the jumping-to-conclusions bias and the bias against disconfirmatory evidence seen in active delusions (Dudley et al., 2016; McLean, Mattiske, & Balzan, 2017). Neutral outputs are further

personalized through aberrant salience (Başaran & Coşar, 2026; Corlett & Fraser, 2025; Kapur, 2003).

Empirically, the SIM-VAIL auditing framework tested 9 frontier chatbots and found concerning behavior accumulated over conversational turns, highest when otherwise supportive behaviors reinforced psychological vulnerabilities (Weilnhammer et al., 2026). Consistent with this, a *JAMA Psychiatry* evaluation found all tested ChatGPT versions produced high rates of inappropriate responses to psychotic prompts, with the free tier performing worst (Shen et al., 2026b).

[Figure 3: Sycophancy of friendly LLMs increases when users express sadness (Ong, 2026; Ibrahim et al., 2026) — see web version for chart.]

Vulnerable Populations

The literature points to overlapping high-risk groups:

- Individuals at clinical high risk / prodromal for psychosis (Buck & Maheux, 2026).
- Those with established psychotic or bipolar disorders (Hudon & Pagé, 2026; Kállai et al., 2026).
- People in substance-induced states (Shah & Morrin, 2026).
- The socially isolated and lonely (Dohnány et al., 2026).
- Adolescents and young adults (Diaz et al., 2026; Hinduja & Patchin, 2026).
- The economically disadvantaged reliant on free tiers (Shen et al., 2026b).

Secondary mechanisms, such as nighttime use causing sleep deprivation and progressive social withdrawal, likely compound this risk (D'Ambrosio et al., 2026; Johnson, Bower, & Reeve, 2026; Treuer & Incze, 2026).

Clinical Management

No formal guidelines exist; recommendations are drawn from case reports and expert consensus. Emerging practice points include:

- **Screen routinely** for chatbot use as part of psychiatric assessment, and characterize the LLM's role (catalyst, amplifier, coauthor, or object) (Başaran & Coşar, 2026; Flathers, Roux, & Torous, 2026; Shen, Girgis, & Jutla, 2026; Shen et al., 2026b).
- **Maintain antipsychotic continuity**, since chatbot validation may prompt self-discontinuation (Başaran & Coşar, 2026; Shah & Morrin, 2026).
- **Environmental containment** by restricting AI access during acute episodes (Shah & Morrin, 2026).

- **Psychoeducation** framing LLM outputs as statistical text generation, not intentional communication (Başaran & Coşar, 2026; Shah & Morrin, 2026).
- **"AI-informed care" framework:** personalized instruction protocols, reflective check-ins, and digital advance statements reframing the AI as an "epistemic ally" (Morrin et al., 2026).

Regulatory and Professional Response

The regulatory landscape is fragmented and lagging. In the US, the FDA's authority is constrained by the exclusion of general-wellness software, while the FTC issued orders to companies seeking data on chatbot harms to young people (Angus et al., 2025; Perlis, 2026b; Shim, 2026). State action is outpacing federal (e.g., California's SB 243). Internationally, the EU AI Act mandates disclosure that users are not talking to a human, and the WHO has issued responsible-use guidance (Shim, 2026).

Professional voices are increasingly vocal — Allen Frances warned that chatbots are "dangerous for the minority who have more severe problems" (Frances, 2026). Meanwhile, only 16% of LLM chatbot studies underwent clinical efficacy testing, and harms reporting in RCTs is largely uninterpretable (Hua et al., 2025; Yang et al., 2026).

Evidence Gaps and the Bottom Line

The single most important limitation is the absence of any longitudinal or interventional data. The entire evidence base consists of case reports, surveys, and chatbot-auditing studies, which cannot establish causality or distinguish precipitation from reverse causation (Buck & Maheux, 2026; Morrin et al., 2026). Reporting bias likely inflates apparent severity, the denominator of uneventful chatbot use is unknown, and rapid model turnover makes product-specific safety assessments quickly obsolete (Rubin, 2026; Shen et al., 2026b).

On balance, the current evidence supports treating AI-associated psychosis as a real, mechanistically plausible clinical signal that most convincingly acts as an **amplifier** or **object** of psychosis in vulnerable individuals. Its capacity to catalyze de novo psychosis in healthy people remains unproven.

References

- Angus, D. C., Khera, R., Lieu, T., et al. (2025). AI, health, and health care today and tomorrow. *JAMA*, 334(18), 1650–1664. https://doi.org/10.1001/jama.2025.18490

- Başaran, A. S., & Coşar, B. (2026). Psychotic episode concurrent with interaction with a large language model (LLM): A case report. *BMC Psychiatry*, 26(1), 642. https://doi.org/10.1186/s12888-026-08288-3
- Buck, B., & Maheux, A. J. (2026). Psychosis risk and generative artificial intelligence use frequency, motivations, and delusion-like experiences: Cross-sectional survey study. *Journal of Medical Internet Research*, 28, e85038. https://doi.org/10.2196/85038
- Burns, A. V., Nelson, K., Wang, H., Hegarty, E. M., & Cohn, A. B. (2025). 'The algorithm is hacked': Analysis of technology delusions in a modern-day cohort. *The British Journal of Psychiatry*, 1–5. https://doi.org/10.1192/bjp.2025.10452
- Cheng, M., Lee, C., Khadpe, P., et al. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. *Science*, 391(6792), eaec8352. https://doi.org/10.1126/science.aec8352
- Compton, M. T. (2003). Internet delusions. *Southern Medical Journal*, 96(1), 61–63. https://doi.org/10.1097/01.SMJ.0000047722.98207.32
- Corlett, P. R., & Fraser, K. M. (2025). 20 years of aberrant salience in psychosis: What have we learned? *The American Journal of Psychiatry*, 182(9), 819–829. https://doi.org/10.1176/appi.ajp.20240556
- D'Ambrosio, S., Cavallotti, S., Del Giudice, R., Dibenedetto, C., & D'Agostino, A. (2026). Sleep and circadian disruption as transdiagnostic drivers of psychosis: A unified mechanistic framework. *Expert Review of Neurotherapeutics*, 1–20. https://doi.org/10.1080/14737175.2026.2718395
- Diaz, A. D., Leong, A. W., Bilge-Johnson, S., et al. (2026). Clinical assessment and implications of parasocial relationships in adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry*, 65(3), 331–336. https://doi.org/10.1016/j.jaac.2025.07.005
- Dohnány, S., Kurth-Nelson, Z., Spens, E., et al. (2026). Technological folie à deux: Feedback loops between AI chatbots and mental health. *Nature Mental Health*, 4(3), 336–345. https://doi.org/10.1038/s44220-026-00595-8
- Dudley, R., Taylor, P., Wickham, S., & Hutton, P. (2016). Psychosis, delusions and the "jumping to conclusions" reasoning bias: A systematic review and meta-analysis. *Schizophrenia Bulletin*, 42(3), 652–665. https://doi.org/10.1093/schbul/sbv150
- Flathers, M., Roux, S., & Torous, J. (2026). Beyond artificial intelligence psychosis: A functional typology of large language model-associated psychotic phenomena. *The Lancet Digital Health*, 8(4), 100974. https://doi.org/10.1016/j.landig.2025.100974

- Frances, A. (2026). Warning: AI chatbots will soon dominate psychotherapy. *The British Journal of Psychiatry*, 228(5), 474–478. https://doi.org/10.1192/bjp.2025.10380
- Hinduja, S., & Patchin, J. W. (2026). Risks and harms of conversational artificial intelligence (CAI) chatbot use among US youth. *Journal of Adolescence*, 98(5), 1597–1606. https://doi.org/10.1002/jad.70164
- Hirjak, D., & Fuchs, T. (2010). Delusions of technical alien control: A phenomenological description of three cases. *Psychopathology*, 43(2), 96–103. https://doi.org/10.1159/000274178
- Hua, Y., Siddals, S., Ma, Z., et al. (2025). Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: A systematic review. *World Psychiatry*, 24(3), 383–394. https://doi.org/10.1002/wps.21352
- Hudon, A., & Pagé, C. (2026). Artificial intelligence centrality in psychotic delusions and violence risk in forensic psychiatry: Retrospective observational study of judicial decisions. *Journal of Medical Internet Research*, 28, e93349. https://doi.org/10.2196/93349
- Ibrahim, L., Hafner, F. S., & Rocher, L. (2026). Training language models to be warm can reduce accuracy and increase sycophancy. *Nature*, 652(8112), 1159–1165. https://doi.org/10.1038/s41586-026-10410-0
- Ittefaq, M., Arif, R., & Kamboh, S. A. (2026). Suicide and self-harm in the age of generative artificial intelligence: The case of ChatGPT and rise of suicide-related lawsuits. *Journal of Affective Disorders*, 411, 122025. https://doi.org/10.1016/j.jad.2026.122025
- Johnson, M. R., Bower, J. L., & Reeve, S. (2026). Testing the mediating role of negative affect, affect intolerance, and dissociation in the relationship between sleep loss and psychotic experiences. *Journal of Psychiatric Research*, 201, 207–214. https://doi.org/10.1016/j.jpsychires.2026.06.028
- Kállai, J., Rózsa, S., Bandi, S., et al. (2026). Attitudes toward artificial intelligence among individuals with anxiety, depression, bipolar disorders, and schizophrenia. *Journal of Psychiatric Research*, 198, 224–233. https://doi.org/10.1016/j.jpsychires.2026.03.028
- Kapur, S. (2003). Psychosis as a state of aberrant salience: A framework linking biology, phenomenology, and pharmacology in schizophrenia. *The American Journal of Psychiatry*, 160(1), 13–23. https://doi.org/10.1176/appi.ajp.160.1.13
- McLean, B. F., Mattiske, J. K., & Balzan, R. P. (2017). Association of the jumping to conclusions and evidence integration biases with delusions in

psychosis: A detailed meta-analysis. *Schizophrenia Bulletin*, 43(2), 344–354. https://doi.org/10.1093/schbul/sbw056

- Morrin, H., Nicholls, L., Levin, M., et al. (2026). Artificial intelligence-associated delusions and large language models: Risks, mechanisms of delusion co-creation, and safeguarding strategies. *The Lancet Psychiatry*, 13(6), 522–530. https://doi.org/10.1016/S2215-0366(25)00396-7
- Olsen, S. G., Reinecke-Tellefsen, C. J., & Østergaard, S. D. (2026). Potentially harmful consequences of artificial intelligence (AI) chatbot use among patients with mental illness: Early data from a large psychiatric service system. *Acta Psychiatrica Scandinavica*, 153(4), 301–303. https://doi.org/10.1111/acps.70068
- Ong, D. (2026). Friendlier LLMs tell users what they want to hear — even when it is wrong. *Nature*, 652(8112), 1134–1135. https://doi.org/10.1038/d41586-026-01153-z
- Perlis, R. H., Gunning, F. M., Usla, A., et al. (2026a). Generative AI use and depressive symptoms among US adults. *JAMA Network Open*, 9(1), e2554820. https://doi.org/10.1001/jamanetworkopen.2025.54820
- Perlis, R. H. (2026b). Artificial intelligence and the potential transformation of mental health. *JAMA Psychiatry*. https://doi.org/10.1001/jamapsychiatry.2025.4116
- Rubin, R. (2026). Millions turn to AI chatbots for mental health support. *JAMA*. https://doi.org/10.1001/jama.2025.23965
- Santos, B. D. S., Roza, T. H., & Passos, I. C. (2026). The engagement-validation loop: Sycophancy, anthropomorphic projection, and clinical risk in informal AI-based mental health support. *Journal of Affective Disorders*, 412, 122123. https://doi.org/10.1016/j.jad.2026.122123
- Shah, S., & Morrin, H. (2026). Substance-induced manic psychosis in which delusions were corroborated by a chatbot — Case report. *BMC Psychiatry*. https://doi.org/10.1186/s12888-026-08137-3
- Shen, E., Girgis, R. R., & Jutla, A. (2026a). Psychiatric risk associated with large language model chatbots: An emerging concern. *Psychiatric Services*, 77(7), 681–684. https://doi.org/10.1176/appi.ps.20250597
- Shen, E., Hamati, F., Donohue, M. R., et al. (2026b). Evaluation of large language model chatbot responses to psychotic prompts. *JAMA Psychiatry*. https://doi.org/10.1001/jamapsychiatry.2026.0249
- Shim, R. S. (2026). Why AI's mental health risks demand regulation now. *JAMA Health Forum*, 7(7), e263056. https://doi.org/10.1001/jamahealthforum.2026.3056

- Skodlar, B., Dernovsek, M. Z., & Kocmur, M. (2008). Psychopathology of schizophrenia in Ljubljana (Slovenia) from 1881 to 2000. *The International Journal of Social Psychiatry*, 54(2), 101–111. https://doi.org/10.1177/0020764007083875
- Stompe, T., Ortwein-Swoboda, G., Ritter, K., & Schanda, H. (2003). Old wine in new bottles? Stability and plasticity of the contents of schizophrenic delusions. *Psychopathology*, 36(1), 6–12. https://doi.org/10.1159/000069658
- Treuer, T., & Incze, A. (2026). Chatbot relational dependence and psychosis risk: Conversational artificial intelligence as a potential behavioral addiction context. *Journal of Behavioral Addictions*, 15(2), 533–539. https://doi.org/10.1556/2006.2026.00080
- Vasiliu, O. (2026). Problematic use of generative artificial intelligence chatbots: Current stage of conceptual and clinical understanding. *Frontiers in Public Health*, 14, 1816917. https://doi.org/10.3389/fpubh.2026.1816917
- Weilnhammer, V., Hou, K. Y., Luettgau, L., et al. (2026). A clinically validated framework for auditing AI chatbot behavior in mental health interactions. *Nature Medicine*. https://doi.org/10.1038/s41591-026-04577-2
- Yang, H., Chang, F., Muroi, F., et al. (2026). Commercial AI-based mental health chatbots as low-intensity adjuncts to psychotherapy: Effectiveness, adherence, and safety — A systematic review and meta-analysis. *Psychotherapy and Psychosomatics*, 1–24. https://doi.org/10.1159/000552072