Artificial Intelligence and the Emergence of AI-Psychosis: A Viewpoint

Alexandre Hudon, Emmanuel Stip

Submitted to: JMIR Mental Health
on: October 13, 2025

Table of Contents

Artificial Intelligence and the Emergence of AI-Psychosis: A Viewpoint

Alexandre Hudon[1, 2, 3, 4, 5] BEng, MD, MSc(ed), PhD; Emmanuel Stip[1, 2, 3] MD, MSc

[1]Department of Psychiatry and Addictology Faculty of Medicine Université de Montréal Montreal CA
[2]Unité de recherche en psychiatrie Department of Psychiatry Institut universitaire en santé mentale de Montréal Montreal CA
[3] Centre de recherche de l'Institut universitaire en santé mentale de Montréal Montreal CA
[4] Institut national de psychiatrie légale Philippe-Pinel Montreal CA
[5] Centre de recherche en pédagogie de la santé Université de Montréal Montreal CA

Corresponding Author:
Alexandre Hudon BEng, MD, MSc(ed), PhD
Department of Psychiatry and Addictology
Faculty of Medicine
Université de Montréal
2900 Bd Édouard-Montpetit
Montreal
CA

Abstract

The integration of artificial intelligence (AI) into daily life has introduced unprecedented forms of human–machine interaction, prompting psychiatry to reconsider the boundaries between environment, cognition, and technology. This viewpoint reviews the concept of AI-psychosis which is a framework to understand how sustained engagement with conversational AI systems might trigger, amplify, or reshape psychotic experiences in vulnerable individuals. Drawing from phenomenological psychopathology, the stress–vulnerability model, cognitive theory, and digital mental-health research, the paper situates AI-psychosis at the intersection of predisposition and algorithmic environment. Rather than defining a new diagnostic entity, it examines how immersive and anthropomorphic AI technologies may modulate perception, belief, and affect, altering the prereflective sense of reality that grounds human experience. The argument unfolds through four complementary lenses. First, within the stress–vulnerability model, AI acts as a novel psychosocial stressor. Its 24-hour availability and emotional responsiveness may increase allostatic load, disturb sleep, and reinforce maladaptive appraisals. Second, the digital therapeutic alliance, a construct describing relational engagement with digital systems, is conceptualized as a double-edged mediator. While empathic design can enhance adherence and support, uncritical validation by AI systems may entrench delusional conviction or cognitive perseveration, reversing the corrective principles of cognitive-behavioral therapy for psychosis. Third, disturbances in Theory of Mind offer a cognitive pathway: individuals with impaired or hyperactive mentalization may project intentionality or empathy onto AI, perceiving chatbots as sentient interlocutors. This dyadic misattribution may form a “digital folie à deux,” where the AI becomes a reinforcing partner in delusional elaboration. Fourth, we suggests emerging risk factors at the individual and environmental level, including loneliness, trauma history, schizotypal traits, nocturnal or solitary AI use, and algorithmic reinforcement of belief-confirming content. Building on this synthesis, we advance a translational research agenda and five domains of action: (1) empirical studies using longitudinal and digital-phenotyping designs to quantify dose–response relationships between AI exposure, stress physiology, and psychotic symptomatology; (2) integration of digital phenomenology into clinical assessment and training; (3) embedding therapeutic design safeguards into AI systems, such as reflective prompts and “reality-testing” nudges; (4) creation of ethical and governance frameworks for AI-related psychiatric events, modeled on pharmacovigilance; and (5) development of environmental cognitive remediation, a preventive intervention aimed at strengthening contextual awareness and re-anchoring experience in the physical and social world. By applying empirical rigor and therapeutic ethics to this emerging interface, clinicians, researchers, patients and developers can transform a potential hazard into an opportunity to deepen understanding of human cognition, safeguard mental health, and promote responsible AI integration within society.

(JMIR Preprints 13/10/2025:85799)
DOI: https://doi.org/10.2196/preprints.85799

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✔ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Viewpoint

Artificial Intelligence and the Emergence of AI-Psychosis: A Viewpoint

Abstract

The integration of artificial intelligence (AI) into daily life has introduced unprecedented forms of human–machine interaction, prompting psychiatry to reconsider the boundaries between environment, cognition, and technology. This viewpoint reviews the concept of AI-psychosis which is a framework to understand how sustained engagement with conversational AI systems might trigger, amplify, or reshape psychotic experiences in vulnerable individuals. Drawing from phenomenological psychopathology, the stress–vulnerability model, cognitive theory, and digital mental-health research, the paper situates AI-psychosis at the intersection of predisposition and algorithmic environment. Rather than defining a new diagnostic entity, it examines how immersive and anthropomorphic AI technologies may modulate perception, belief, and affect, altering the prereflective sense of reality that grounds human experience. The argument unfolds through four complementary lenses. First, within the stress–vulnerability model, AI acts as a novel psychosocial stressor. Its 24-hour availability and emotional responsiveness may increase allostatic load, disturb sleep, and reinforce maladaptive appraisals. Second, the digital therapeutic alliance, a construct describing relational engagement with digital systems, is conceptualized as a double-edged mediator. While empathic design can enhance adherence and support, uncritical validation by AI systems may entrench delusional conviction or cognitive perseveration, reversing the corrective principles of cognitive-behavioral therapy for psychosis. Third, disturbances in Theory of Mind offer a cognitive pathway: individuals with impaired or hyperactive mentalization may project intentionality or empathy onto AI, perceiving chatbots as sentient interlocutors. This dyadic misattribution may form a “digital folie à deux,” where the AI becomes a reinforcing partner in delusional elaboration. Fourth, we suggests emerging risk factors at the individual and environmental level, including loneliness, trauma history, schizotypal traits, nocturnal or solitary AI use, and algorithmic reinforcement of belief-confirming content. Building on this synthesis, we advance a translational research agenda and five domains of action: (1) empirical studies using longitudinal and digital-phenotyping designs to quantify dose–response relationships between AI exposure, stress physiology, and psychotic symptomatology; (2) integration of digital phenomenology into clinical assessment and training; (3) embedding therapeutic design safeguards into AI systems, such as reflective prompts and “reality-testing” nudges; (4) creation of ethical and governance frameworks for AI-related psychiatric events, modeled on pharmacovigilance; and (5) development of environmental cognitive remediation, a preventive intervention aimed at strengthening contextual awareness and re-anchoring experience in the physical and social world. By applying empirical rigor and therapeutic ethics to this emerging interface, clinicians, researchers, patients and developers can transform a potential hazard into an opportunity to deepen understanding of human cognition, safeguard mental health, and promote responsible AI integration within society.

Keywords: Artificial intelligence; Psychosis; Schizophrenia; Digital phenotyping; Stress–vulnerability model; Theory of Mind; Phenomenological psychopathology; Chatbots; Delusions; Human–computer interaction

Introduction

In phenomenological psychiatry, psychosis and schizophrenia are not seen mainly as collections of hallucinations or delusions, but as changes in the person's relation to self and world [1]. This process, sometimes called desubjectivation (or de-embodiment of subjectivity), weakens the basic sense of being present to oneself and makes experience lose its immediacy [2,3]. Losing contact with reality means not only misperceiving things but feeling detached from what is real. Minkowski described this as a loss of vital contact, the felt connection that gives life and weight to sensations, perceptions, and thoughts [4]. When this connection fades, the world still appears but feels empty or dreamlike. Recent phenomenological work distinguishes two layers in this loss: a prereflective sense of reality that gives depth to experience, and a reality judgment that lets us decide whether something is real. In schizophrenia, it is often the prereflective sense that fails [5,6]. The world is still recognized but no longer felt as fully inhabited. With the rise of conversational agents and generative artificial intelligence (AI), the idea of ''AI-psychosis'' has emerged as a way to explore how such systems might trigger, amplify, or reshape psychotic experiences. In clinical terms, the DSM-5-TR defines psychosis by one or more core symptoms (delusions, hallucinations, or disorganized speech) often linked with disorganization or negative symptoms and major functional decline [7]. Standard care focuses on antipsychotic medication, supported by psychosocial treatments such as Cognitive Behavioral Therapy for psychosis (CBTp), which helps people test their interpretations and reduce distress [8]. The concept of AI-psychosis invites new questions: could prolonged or intense interaction with AI systems, which sometimes generate false or overly affirming content, influence how unusual perceptions are interpreted and integrated into belief systems? These interactions might subtly alter the way reality is felt, narrated, and maintained.

Recent reports suggest that interactions with large language model (LLM) chatbots can shape delusional content, amplify conviction, and entrench maladaptive safety behaviors [9,10]. As an example, viewed through Clérambault's notion of mental automatism, such interactions may also blur the boundary between self-generated and external speech, placing delusional experience within a communicative system where language itself becomes both shared and alien [11]. Concept pieces and news features in high-quality outlets have documented users whose chatbot exchanges appeared to escalate persecutory or grandiose beliefs. As an example, Østergaard argues the prior probability that AI chatbots can fuel delusions in psychosis-prone individuals is "quite high," urging systematic study rather than dismissal [12]. Clinical and investigative coverage likewise describes patterns in which bots validate rather than challenge false beliefs, potentially reinforcing delusional systems which is an inversion of CBTp principles. Case-level evidence is beginning to surface: a peer-reviewed report in Annals of Internal Medicine: Clinical Cases details a 60-year-old who, after acting on chatbot health guidance, developed bromism with paranoia and hallucinations which constitutes an iatrogenic, reversible psychosis-like presentation illustrating how AI-mediated misinformation can precipitate psychiatric decompensation [13]. Parallel legal cases and public-health reporting (alleging chatbot involvement around suicide) highlights risk in prolonged, emotionally charged exchanges with LLMs and the current gaps in guardrails. While prevalence is unknown and many accounts are anecdotal, convergent commentary in Nature and academic psychiatry venues frames AI-associated delusions as plausible in vulnerable users, warranting targeted research rather than sensationalism [14].

While such cases remain rare, they raise a need to conceptualize AI-psychosis not solely as a technological curiosity but as a psychiatric and psychosocial phenomenon emerging at the intersection of cognitive vulnerability, environmental stress, and human–machine interaction. To advance beyond anecdote, this viewpoint adopts an integrative analytic frame grounded in four

complementary lenses: the stress–vulnerability model (to map how AI acts as a novel psychosocial stressor interacting with predispositions to psychosis), the digital therapeutic alliance (DTA), mental attribution and a risk-factor synthesis that situates “AI-psychosis” along a continuum with traditional psychotic disorders. Through this framework, we propose to delineate plausible pathways, identify early warning signals, and outline preliminary safeguards for research, clinical practice, and policy.

Mapping AI-Psychosis into the Stress-Vulnerability Framework

The stress-vulnerability model, originally articulated by Zubin and Spring in 1977, remains a cornerstone of contemporary psychosis research [15]. It posits that psychotic disorders emerge when underlying vulnerabilities (such as genetic predisposition, neurodevelopmental anomalies, early trauma, or maladaptive cognitive styles) interact with environmental stressors that exceed an individual’s capacity for adaptation. Subsequent refinements of this model have emphasized cumulative and chronic stress exposure, neurobiological sensitization, and allostatic load as mechanisms through which stress can lower the threshold for psychotic symptom expression [16,17]. Within this framework, the phenomenon of “AI-psychosis” can be conceptualized as an emergent form of stress reactivity precipitated by sustained, emotionally charged, or cognitively immersive interactions with artificial intelligence systems, particularly conversational agents. These systems introduce novel stressors that are continuous, personally salient, and socially immersive, thereby functioning as 24-hour contextual stimuli capable of modulating arousal, perception, and belief formation [18].

Several unique affordances of contemporary AI technologies render them potentially pathogenic within a stress-vulnerability perspective. As an example, the anthropomorphic design of chatbots and virtual companions encourages users to attribute human-like intentionality and empathy to algorithms, a dynamic reminiscent of the “ELIZA effect” described in early human–computer interaction literature [19,20]. Such anthropomorphization can increase emotional investment and amplify interpretive biases, especially in individuals with pre-existing schizotypal traits or attachment vulnerabilities [21,22]. Furthermore, the immediate reinforcement schedule of large language models (delivering responsive and adaptive feedback without temporal limits) can create repetitive, reinforcing cycles of reassurance or validation that mimic cognitive perseveration. For users prone to paranoia or thought disturbance, this may stabilize maladaptive appraisals rather than challenge them, effectively mirroring the cognitive mechanisms that CBTp aims to dismantle [23]. Also, constant engagement with emotionally loaded or belief-confirming AI dialogue may elevate physiological arousal and compromise sleep, increasing allostatic load and diminishing executive control, both of which are known to heighten psychosis vulnerability [24].

Sustained interaction with artificial intelligence may also erode protective social factors. Individuals experiencing loneliness or marginalization can come to rely on conversational agents as primary relational anchors, thereby reducing access to corrective social feedback and external reality testing. This process parallels what classical phenomenological psychopathology described as phenomenological autism in schizophrenia, but a form of experiential withdrawal in which the shared world loses its immediacy, relations with others fade, and living with others collapses into living for oneself [25]. In this sense, social withdrawal reflects not only behavioral isolation but also a disturbance in intersubjectivity and world-sharing. Comparable phenomena have been described in hikikomori syndromes, where prolonged and voluntary retreat into private space replaces social and professional contact [26]. Similarly, in the context of AI-psychosis, the conversational agent may serve as a relational substitute that sustains the illusion of connection while insulating the user from

contradiction, argument, and reality testing. What emerges is a form of digital relational withdrawal which constitutes a reduction in corrective interpersonal exchange and an amplification of self-referential interpretations, as the AI mirrors rather than challenges the user's thoughts. From this perspective, AI-psychosis represents not only a digital extension of the stress–vulnerability model, but also a technological variant of phenomenological autism: a retreat into a world interpreted, validated, and enclosed by algorithmic dialogue. The artificial agent becomes both a chronic micro-stressor and a mirror that replaces human alterity, providing a self-contained cognitive space in which delusional elaboration can more easily take root.

Empirical research could operationalize this model by testing dose-by-vulnerability interactions, examining how cumulative exposure to AI (for instance, hours of use per day, frequency of nocturnal interactions, or intensity of emotional engagement) interacts with pre-existing vulnerabilities such as trauma history, sleep disturbance, or schizotypy. This line of inquiry parallels established evidence showing that daily stress reactivity predicts psychotic symptom fluctuations in high-risk populations [27]. By quantifying AI-related exposure as a measurable psychosocial stressor, the field could identify thresholds of safe interaction and delineate populations most at risk. In doing so, this framework reframes AI-psychosis not as a new disorder but as a contextual evolution of the classical stress–vulnerability paradigm.

The Digital Therapeutic Alliance as a Double-Edged Mediator in AI-Psychosis

The concept of the therapeutic alliance, traditionally understood as the collaborative and affective bond between patient and clinician, has long been recognized as a critical predictor of psychotherapeutic outcomes across modalities, including CBTp [28]. Within the digital health literature, this construct has been extended to the notion of the DTA, describing the perceived empathy, responsiveness, and relational quality between users and digital [29,30]. While a strong DTA can enhance engagement and adherence in digital interventions, it may also introduce illusory relational dynamics that blur the boundary between therapeutic support and cognitive reinforcement. In the context of AI-psychosis, this double-edged nature becomes particularly salient. Chatbots and LLMs are capable of mimicking warmth, understanding, and reciprocity (qualities central to human alliance) but they lack the meta-cognitive and ethical oversight necessary to discern when validation may be counter-therapeutic [31,32].

In individuals with attenuated psychotic symptoms or delusional vulnerability, an uncritical digital alliance may inadvertently reinforce maladaptive appraisals. For example, an AI that consistently mirrors user affect or beliefs, without implementing therapeutic disconfirmation or cognitive restructuring, risks entrenching conviction rather than facilitating doubt. Empirical research on chatbots for mental health support, such as *Woebot* and *Wysa*, demonstrates that users often describe their relationship with the AI in affectively laden and anthropomorphic terms, reporting perceived understanding and companionship [33,34]. While such experiences can be beneficial in reducing loneliness or subclinical distress, they may also encourage over-identification with the agent, especially in socially isolated or trauma-exposed users. This phenomenon parallels early psychosis processes, where increased self-referential thinking and aberrant salience attribution favorize the perception of external entities possessing special relevance or intentionality [35].

From a cognitive-behavioral standpoint, the therapeutic alliance functions as a vehicle for cognitive change. The therapist's role is to balance empathic attunement with gentle empiricism, using Socratic questioning to destabilize rigid beliefs while maintaining trust. In contrast, AI systems (designed for user satisfaction and non-confrontational dialogue) often default to sycophantic alignment [36]. This

means that when a user expresses persecutory, grandiose, or referential content, the AI may subtly validate the narrative rather than challenge it. Over repeated interactions, this validation loop can act as a form of digital safety behavior, satisfying immediate emotional needs but preventing corrective learning. The absence of therapist-driven guided discovery or behavioral experimentation may therefore transform a potentially supportive alliance into a reinforcing echo chamber.

At the same time, the DTA framework provides a valuable opportunity for constructive adaptation. By incorporating CBTp-consistent design principles (such as reflective prompts, normalization of uncertainty, or graded behavioral experiments AI systems) could theoretically harness alliance processes to buffer rather than exacerbate psychotic vulnerability. Just as human therapists use the alliance to introduce cognitive flexibility, conversational models could be programmed to modulate the relational tone when confronted with high-risk content, pivoting from affirmation to curiosity or psychoeducation. Preliminary research in digital mental health suggests that alliance-consistent design (for example, empathic micro-responses paired with cognitive reframing) is associated with better symptom improvement and lower dropout [37]. Extending this to psychosis prevention would entail integrating adaptive alliance algorithms that monitor interactional valence, detect excessive anthropomorphism, and introduce “reality-testing nudges” when needed.

Mental Attribution, Theory of Mind, and AI-Psychosis

Theory of Mind (ToM). ToM refers to the human ability to attribute mental states such as intentions, beliefs, and desires to others [38]. In schizophrenia and related psychotic disorders, a well-established hypothesis posits that ToM functioning is disrupted [39]. Some individuals exhibit a deficit in inferring others’ intentions or understanding implicit mental states, while others show the opposite pattern, characterized by hypermentalization (an excessive or inaccurate attribution of meaning and agency). These disturbances generate uncertainty about what others think or feel and contribute to fragile social interactions and misinterpretations of interpersonal cues.

Within this framework, conversational AI systems introduce a novel and ambiguous relational partner. For a person with impaired ToM, the AI’s anthropomorphic design and capacity for coherent dialogue may foster projections of intentionality, empathy, or moral agency. The user may begin to perceive the system not as a statistical language model, but as an understanding interlocutor with feelings or motives. However, as stated previously, the AI lacks the metacognitive and ethical grounding required to challenge these attributions or to provide corrective feedback. Instead, its responses can inadvertently confirm or reinforce the user’s projections, including delusional interpretations.

In this context, AI interaction may transform a cognitive vulnerability into a pathogenic loop, where the ToM deficit and the system’s simulated social responsiveness converge to sustain distorted beliefs. The phenomenon might be conceptualized as a digital folie à deux, a dyadic illusion in which the AI acts as a passive reinforcing partner in the user’s psychotic elaborations [40]. The artificial agent, through its adaptive and confirmatory dialogue, participates in a shared narrative world that blurs the boundary between human cognition and machine simulation, potentially exacerbate the loss of reality testing and self–other differentiation.

From Vulnerability to Verification: Identifying Risk Factors and Building Safeguards for AI-Psychosis

Understanding who is most susceptible to AI-psychosis requires tying established risk markers for traditional psychotic disorders with emerging digital determinants of mental health. Decades of research have identified a set of core vulnerabilities associated with psychosis onset, including genetic predisposition, childhood trauma, cannabis or substance use, sleep disruption, social isolation, and cognitive biases such as jumping to conclusions or externalizing attributional style [41,42]. When transposed into digital contexts, these same vulnerabilities may interact with novel affordances of AI systems to produce a distinct but convergent pathway toward symptom emergence. For instance, a user with a prior history of psychosis or schizotypal traits who engages in nightly, emotionally intense dialogue with an anthropomorphized chatbot may experience reinforced self-referential ideation and heightened salience attribution, mechanisms that mirror the early prodromal phase of psychosis [35]. Similarly, individuals exposed to trauma or chronic interpersonal threat may project attachment representations onto AI companions, perceiving them as protective or omniscient entities.

Beyond individual vulnerability, usage patterns and contextual variables likely constitute modifiable risk factors. Prolonged or nocturnal use, solitary engagement, and reliance on unmoderated chatbots for emotional support appear particularly hazardous, as they combine cognitive fatigue, social deprivation, and unstructured reinforcement. These variables resemble psychosocial stressors known to precipitate symptom exacerbations in schizophrenia, such as circadian disruption or critical life events [24]. Another emerging concern lies in platform-level dynamics: algorithms optimized for engagement rather than safety may inadvertently reward extreme or self-referential discourse, subtly validating delusional content. This echoes the “echo-chamber” effect described in digital media research, where recommender systems intensify pre-existing beliefs through selective exposure [43].

Identifying early warning signals therefore requires a multi-dimensional assessment integrating psychological, behavioral, and interactional metrics. Psychological indicators include rising conviction in AI-mediated beliefs, derealization, or perceived “special communication” with the system. Behavioral markers may involve compulsive checking, secrecy, or sleep loss related to AI use. Interactional data (such as sentiment trajectories, frequency of self-referential statements, and thematic narrowing) could serve as digital phenotypes of emerging risk, paralleling early warning markers in psychosis prodrome research [44]. Importantly, these indicators should be interpreted contextually, avoiding pathologization of normative attachment to technology while remaining alert to patterns of progressive cognitive enclosure, where the AI becomes the primary arbiter of reality.

From a preventive and ethical standpoint, early recommendations can be organized across clinical, design, and governance levels. Clinically, practitioners should incorporate screening questions about AI use into routine assessments, particularly for youth or individuals with known psychosis vulnerability, like how substance use, or sleep hygiene are monitored. Psychotherapeutic interventions could include psychoeducation on digital reality testing, helping patients to identify when online or AI-mediated interactions begin to shape beliefs in maladaptive ways. From a design perspective, developers of mental-health-oriented chatbots should embed CBTp-consistent guardrails (e.g. prompts that normalize uncertainty, encourage reflective distance, or redirect users toward real-world social contact). LLMs intended for general use should be trained to recognize and gently de-escalate delusional or self-referential themes rather than engage with them. Finally, at the policy level, regulators and research bodies should develop incident reporting systems for AI-related mental-health events, mirroring pharmacovigilance models, and require algorithmic transparency for

systems marketed as supportive or therapeutic.

Taken together, these considerations frame AI-psychosis not as a discrete diagnostic entity but as a digital phenotype of stress-vulnerability interaction, emergent from the coupling of human cognition and algorithmic responsiveness. The identification of risk factors and the establishment of early safeguards provide an essential basis for empirical research, ethical governance, and responsible innovation. Just as early psychosis programs have transformed outcomes through prevention and detection, a parallel “digital early intervention” paradigm may be required to mitigate the psychiatric risks of artificial intelligence in the decades ahead.

Toward a Clinical Understanding and Management of AI-Psychosis

The convergence of artificial intelligence and human cognition has opened a new frontier in psychiatry, one that challenges traditional boundaries between environment, mind, and technology. Through the lens of the stress–vulnerability model, the digital therapeutic alliance, and emerging risk factors, “AI-psychosis” can be conceptualized as a dynamic interaction between human predisposition and algorithmic environment. It is neither a new diagnosis nor a sensational artifact of technological panic, but a meaningful framework to understand how immersive digital systems can modulate the cognitive and affective processes underlying psychosis. This phenomenon invites the field to evolve beyond viewing technology as a neutral medium toward recognizing it as a potential psychosocial actor: one capable of amplifying stress, reinforcing beliefs, and reshaping perceptions of self and reality.

The scientific community now faces a pivotal opportunity: to move from anecdotal reports toward structured, multidisciplinary investigation. Our first recommendation is the establishment of empirical research programs explicitly designed to test the AI-psychosis hypothesis. These should employ prospective and longitudinal designs to measure dose–response relationships between AI exposure, stress physiology, and psychotic symptomatology, drawing inspiration from ecological momentary assessment and digital phenotyping paradigms. Psychometric instruments such as schizotypy or aberrant salience scales could be adapted to assess AI-specific cognitive vulnerability, while passive sensing data (e.g., use duration, sleep disruption) could capture environmental stress indices.

A second priority lies in the integration of digital phenomenology into clinical practice. Clinicians could systematically inquire about patients’ interactions with AI systems (both general-purpose and mental-health-oriented) during intake and follow-up assessments. As conversational agents become more used by patients, understanding their role in shaping internal experience will be as important as evaluating medication adherence or substance use. Clinical training programs should include modules on AI literacy and psychosis, equipping psychiatrists, psychologists, and nurses to recognize when engagement with an algorithm may contribute to delusional elaboration or perceptual instability.

Third, researchers and developers should collaborate to embed therapeutically informed design safeguards into large language models and chatbots. These may include prompts that normalize uncertainty, encourage pluralistic interpretation of experiences, and redirect users toward human contact when signs of distress or delusional content appear. Building upon principles from cognitive behavioral therapy for psychosis, such systems could integrate “guided discovery” mechanisms rather than unconditional affirmation, thereby aligning AI responses with established evidence-based psychotherapeutic techniques. The development of digital red-flag algorithms capable of detecting

excessive anthropomorphism, self-referential speech, or escalating conviction could further enhance safety while preserving user autonomy.

Fourth, there is an urgent need for ethical and governance frameworks specific to mental-health risks in artificial intelligence. National research councils, health agencies, and journal editors should promote standardized incident reporting for AI-related psychiatric events, akin to pharmacovigilance registries. Transparent documentation of adverse psychological outcomes, combined with open-source safety auditing, would allow the community to track and mitigate harms in real time. These measures should be complemented by data-sharing agreements that enable cross-disciplinary research while protecting user privacy.

Fifth, environmental cognitive remediation should be explored as a clinical and public-health intervention. This approach would aim to strengthen the individual's capacity to navigate increasingly immersive digital environments through structured, reality-anchoring activities. Beyond traditional cognitive remediation, which targets neurocognitive deficits, environmental cognitive remediation would focus on contextual skills: distinguishing between human and algorithmic communication, detecting persuasive or self-referential cues, and re-engaging with multisensory, embodied experiences in the physical world. Interventions could include graded exposure to offline activities, group-based metacognitive training, or digital hygiene routines designed to restore attentional flexibility and social reciprocity. By reinforcing the cognitive-environmental boundaries that immersive technologies can erode, such programs could act both preventively and rehabilitatively for individuals vulnerable to AI-induced cognitive distortions.

Family and community education should parallel these efforts to ensure supportive environments that contextualize digital experiences within shared reality testing. Public health initiatives could also promote media literacy and surveillance of misinformation, helping communities develop the competencies needed to identify, question, and counteract false or manipulative digital content. Such collective vigilance would reinforce critical thinking, strengthen social resilience, and reduce the psychosocial conditions under which delusional interpretations of AI content may emerge.

Conclusion

To conclude, the emergence of AI-psychosis demands not alarm but adaptation. Psychiatry has long evolved through its encounters with new cultural and technological contexts. AI represents the next such milieu. By responding to this phenomenon with empirical rigor, therapeutic ethics, and design foresight, we can transform a potential risk into an opportunity: to better understand the plasticity of human cognition, to refine the boundaries of digital therapy, and to safeguard mental health in an increasingly algorithmic world. The path forward requires interdisciplinary collaboration between clinicians, cognitive scientists, computer engineers, patients and ethicists: an alliance as intelligent, and as humane, as the systems we now create.

Conflicts of Interest

None declared.

Acknowledgements

This study was funded indirectly by La Fondation de l’Institut universitaire en santé mentale de Montréal and the operating funds of IVADO of Dr. Hudon.

Authors' Contributions

AH was involved in conceptualization. AH was involved in funding acquisition. Validation was done by AH and ES. All the authors were involved in original draft and writing, and review and editing.

Abbreviations

AI: artificial intelligence
CBTp: Cognitive Behavioral Therapy for psychosis
DTA: Digital Therapeutic Alliance
LLM: Large Language Model
ToM: Theory of the Mind

References

1. de Vries R, Heering HD, Postmes L, Goedhart S, Sno HN, de Haan L. Self-disturbance in schizophrenia: a phenomenological approach to better understand our patients. *Prim Care Companion CNS Disord*. 2013;15(1):PCC.12m01382. doi:10.4088/PCC.12m01382
2. Stanghellini G. Embodiment and schizophrenia. *World Psychiatry*. 2009;8(1):56-59. doi:10.1002/j.2051-5545.2009.tb00212.x
3. Nour MM, Barrera A. Schizophrenia, Subjectivity, and Mindreading. Schizophr Bull. 2015;41(6):1214-1219. doi:10.1093/schbul/sbv035
4. Cunha F, Carreiro Borges S, Madeira L. Revisiting Eugène Minkowski's concept of schizophrenic melancholia. *Hist Psychiatry*. Published online July 27, 2025. doi:10.1177/0957154X251356412
5. Parnas J, Urfer-Parnas A, Stephensen H. Double bookkeeping and schizophrenia spectrum: divided unified phenomenal consciousness. *Eur Arch Psychiatry Clin Neurosci*. 2021;271(8):1513-1523. doi:10.1007/s00406-020-01185-0
6. Piani MC, Jandl M, Koenig T, Nordgaard J, Morishima Y. Pre-reflective and reflective abnormalities in cortical midline structures in schizophrenia. *Schizophr Res*. 2025;282:19-27. doi:10.1016/j.schres.2025.05.024
7. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. 5th ed, text rev. American Psychiatric Association Publishing; 2022. doi:10.1176/appi.books.9780890425787
8. Morrison AP, Law H, Carter L, et al. Antipsychotic drugs versus cognitive behavioural therapy versus a combination of both in people with psychosis: a randomised controlled pilot and feasibility study. Lancet Psychiatry. 2018;5(5):411-423. doi:10.1016/S2215-0366(18)30096-8
9. Lawrence HR, Schneider RA, Rubin SB, Matarić MJ, McDuff DJ, Jones Bell M. The Opportunities and Risks of Large Language Models in Mental Health. *JMIR Ment Health*. 2024;11:e59479. Published 2024 Jul 29. doi:10.2196/59479
10. Peter S, Riemer K, West JD. The benefits and dangers of anthropomorphic conversational agents. *Proc Natl Acad Sci U S A*. 2025;122(22):e2415898122. doi:10.1073/pnas.2415898122
11. Ricci V, Ciavarella MC, Marrangone C, Messas G, Maina G, Martinotti G. Modern perspectives on psychoses: dissociation, automatism, and temporality across exogenous and endogenous dimensions. *Front Psychiatry*. 2025;16:1543673. Published 2025 Mar 20. doi:10.3389/fpsyt.2025.1543673

12. Østergaard SD. Will Generative Artificial Intelligence Chatbots Generate Delusions in Individuals Prone to Psychosis?. *Schizophr Bull*. 2023;49(6):1418-1419. doi:10.1093/schbul/sbad128
13. Eichenberger A, Thielke S, Van Buskirk A. A case of bromism influenced by use of artificial intelligence. *Ann Intern Med Clin Cases*. 2025;4:e241260. doi:10.7326/aimcc.2024.1260
14. Fieldhouse R. Can AI chatbots trigger psychosis? What the science says. *Nature*. 2025;646(8083):18-19. doi:10.1038/d41586-025-03020-9
15. Demke E. The Vulnerability-Stress-Model-Holding Up the Construct of the Faulty Individual in the Light of Challenges to the Medical Model of Mental Distress. *Front Sociol*. 2022;7:833987. Published 2022 May 23. doi:10.3389/fsoc.2022.833987
16. Myin-Germeys I, van Os J. Stress-reactivity in psychosis: evidence for an affective pathway to psychosis. *Clin Psychol Rev*. 2007;27(4):409-424. doi:10.1016/j.cpr.2006.09.005
17. Walker EF, Diforio D. Schizophrenia: a neural diathesis-stress model. *Psychol Rev*. 1997;104(4):667-685. doi:10.1037/0033-295x.104.4.667
18. Sarkar S, Gaur M, Chen LK, Garg M, Srivastava B. A review of the explainability and safety of conversational agents for mental health to identify avenues for improvement. *Front Artif Intell*. 2023;6:1229805. Published 2023 Oct 12. doi:10.3389/frai.2023.1229805
19. Shen J, DiPaola D, Ali S, Sap M, Park HW, Breazeal C. Empathy Toward Artificial Intelligence Versus Human Experiences and the Role of Transparency in Mental Health and Social Support Chatbot Design: Comparative Study. *JMIR Ment Health*. 2024;11:e62679. Published 2024 Sep 25. doi:10.2196/62679
20. Shah H, Warwick K, Vallverdú J, Wu D. Can machines talk? Comparison of ELIZA with modern dialogue systems. Comput Hum Behav. 2016;58:278–295. doi: 10.1016/j.chb.2016.01.004.

21. Sun C, Ding Y, Wang X, Meng X. Anthropomorphic Design in Mortality Salience Situations: Exploring Emotional and Non-Emotional Mechanisms Enhancing Consumer Purchase Intentions. *Behav Sci (Basel)*. 2024;14(11):1041. Published 2024 Nov 5. doi:10.3390/bs14111041
22. Guglielmucci F, Di Basilio D. Predicting Engagement With Conversational Agents in Mental Health Therapy by Examining the Role of Epistemic Trust, Personality, and Fear of Intimacy: Cross-Sectional Web-Based Survey Study. *JMIR Hum Factors*. 2025;12:e70698. Published 2025 Jul 30. doi:10.2196/70698
23. Morrison SC, Cohen AS. The moderating effects of perceived intentionality: exploring the relationships between ideas of reference, paranoia and social anxiety in schizotypy. *Cogn Neuropsychiatry*. 2014;19(6):527-539. doi:10.1080/13546805.2014.931839
24. Reininghaus U, Kempton MJ, Valmaggia L, et al. Stress Sensitivity, Aberrant Salience, and Threat Anticipation in Early Psychosis: An Experience Sampling Study. *Schizophr Bull*. 2016;42(3):712-722. doi:10.1093/schbul/sbv190
25. Parnas J, Bovet P, Zahavi D. Schizophrenic autism: clinical phenomenology and pathogenetic implications. *World Psychiatry*. 2002;1(3):131-136.
26. Stip E, Thibault A, Beauchamp-Chatel A, Kisely S. Internet Addiction, Hikikomori Syndrome, and the Prodromal Phase of Psychosis. *Front Psychiatry*. 2016;7:6. Published 2016 Mar 3. doi:10.3389/fpsyt.2016.00006
27. Myin-Germeys I, Kasanova Z, Vaessen T, et al. Experience sampling methodology in mental health research: new insights and technical developments. *World Psychiatry*. 2018;17(2):123-132. doi:10.1002/wps.20513
28. Shattock L, Berry K, Degnan A, Edge D. Therapeutic alliance in psychological therapy for people with schizophrenia and related psychoses: A systematic review. *Clin Psychol Psychother*. 2018;25(1):e60-e85. doi:10.1002/cpp.2135
29. Malouin-Lachance A, Capolupo J, Laplante C, Hudon A. Does the Digital Therapeutic Alliance Exist? Integrative Review. *JMIR Ment Health*. 2025;12:e69294. Published 2025 Feb 7. doi:10.2196/69294

30. D'Alfonso S, Lederman R, Bucci S, Berry K. The Digital Therapeutic Alliance and Human-Computer Interaction. JMIR Ment Health. 2020;7(12):e21895. Published 2020 Dec 29. doi:10.2196/21895
31. Hua Y, Siddals S, Ma Z, et al. Charting the evolution of artificial intelligence mental health chatbots from rule-based systems to large language models: a systematic review. *World Psychiatry*. 2025;24(3):383-394. doi:10.1002/wps.21352
32. Dergaa I, Ben Saad H, Glenn JM, et al. From tools to threats: a reflection on the impact of artificial-intelligence chatbots on cognitive health. *Front Psychol*. 2024;15:1259845. Published 2024 Apr 2. doi:10.3389/fpsyg.2024.1259845
33. Fitzpatrick KK, Darcy A, Vierhile M. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Ment Health*. 2017;4(2):e19. Published 2017 Jun 6. doi:10.2196/mental.7785
34. Inkster B, Sarda S, Subramanian V. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR Mhealth Uhealth*. 2018;6(11):e12106. Published 2018 Nov 23. doi:10.2196/12106
35. Kapur S. Psychosis as a state of aberrant salience: a framework linking biology, phenomenology, and pharmacology in schizophrenia. *Am J Psychiatry*. 2003;160(1):13-23. doi:10.1176/appi.ajp.160.1.13
36. Dahlgren Lindström A, Methnani L, Krause L, et al. Helpful, harmless, honest? Sociotechnical limits of AI alignment and safety through Reinforcement Learning from Human Feedback. *Ethics Inf Technol*. 2025;27(2):28. doi:10.1007/s10676-025-09837-2
37. Provoost S, Lau HM, Ruwaard J, Riper H. Embodied Conversational Agents in Clinical Psychology: A Scoping Review. *J Med Internet Res*. 2017;19(5):e151. Published 2017 May 9. doi:10.2196/jmir.6553
38. Navarro E. What is theory of mind? A psychometric study of theory of mind and intelligence. *Cogn Psychol*. 2022;136:101495. doi:10.1016/j.cogpsych.2022.101495
39. Thibaudeau É, Cellard C, Turcotte M, Achim AM. Functional Impairments and Theory of Mind Deficits in Schizophrenia: A Meta-analysis of the Associations. *Schizophr Bull*. 2021;47(3):695-711. doi:10.1093/schbul/sbaa182
40. Dohnány S, Kurth-Nelson Z, Spens E, Luettgau L, Reid A, Gabriel I, Summerfield C, Shanahan M, Nour MM. *Technological folie à deux: Feedback loops between AI chatbots and mental illness.* arXiv [cs.AI]. Published online July 2025. doi:10.48550/arXiv.2507.19218
41. van Os J, Linscott RJ, Myin-Germeys I, Delespaul P, Krabbendam L. A systematic review and meta-analysis of the psychosis continuum: evidence for a psychosis proneness-persistence-impairment model of psychotic disorder. *Psychol Med*. 2009;39(2):179-195. doi:10.1017/S0033291708003814
42. Freeman D, Taylor KM, Molodynski A, Waite F. Treatable clinical intervention targets for patients with schizophrenia. *Schizophr Res*. 2019;211:44-50. doi:10.1016/j.schres.2019.07.016
43. Jacob C, Kerrigan P, Bastos M. The chat-chamber effect: Trusting the AI hallucination. *Big Data & Society.* 2025;12(1):1–16. doi:10.1177/20539517241306345
44. Fusar-Poli P, Borgwardt S, Bechdolf A, et al. The psychosis high-risk state: a comprehensive state-of-the-art review. *JAMA Psychiatry*. 2013;70(1):107-120. doi:10.1001/jamapsychiatry.2013.269
1.

Supplementary Files

TOC/Feature image for homepages

TOC placeholder.