European Journal of Mental Health
https://doi.org/10.5708/EJMH.21.2026.0066

PERSPECTIVE ARTICLE

This publication is part of the 2026-2027 SPECIAL COMPILATION on "Mental Health in the Digital Era"

Artificial Intelligence and the Possible Transformation of Psychopathology: Amplification, Modification, or Novelty?

Máté SZONDY [1,2] and Máté KAPITÁNY-FÖVÉNY [3,4]

OPEN ACCESS

Affiliations
[1] Semmelweis University, Faculty of Medicine, Department of Clinical Psychology, Budapest, Hungary
[2] Jewish Charity Hospital, Budapest, Hungary
[3] Nyírő Gyula National Institute of Psychiatry and Addictions, Budapest, Hungary
[4] Semmelweis University Faculty of Health Sciences, Department of Addictology, Budapest, Hungary

Correspondence
Máté SZONDY
Semmelweis University, Faculty of Medicine, Department of Clinical Psychology
1085 Budapest Üllői St 26, Hungary
Email: szondy.mate@semmelweis.hu

History
Received: 30 May 2026
Accepted: 12 November 2026
Published: 26 November 2026

Citation
Szondy, M., & Kapitány-Fövény, M. (2026). Artificial intelligence and the possible transformation of psychopathology: Amplification, modification, or novelty? *European Journal of Mental Health, 21*, e0066, 1–11.
https://doi.org/10.5708/EJMH.21.2026.0066

Introduction: Technological environments have long shaped the phenomenology of mental distress by reorganizing perception, cognition, and social interaction. Artificial intelligence (AI), through recommender systems, conversational agents, and synthetic media, may intensify this influence through its interactive, adaptive, and personalized character. We distinguish four conceptually separable ways in which AI may relate to psychopathology: novel diagnostic entities, novel symptom constellations, qualitative (pathoplastic) modification of established disorders, and quantitative amplification of preexisting symptoms or vulnerabilities.
Areas covered: Focusing on the risk side of a broader range of possible effects, which also includes symptom reduction and resilience-enhancing effects, we review six mechanisms through which AI may interact with psychopathological vulnerability: algorithmic amplification, anthropomorphic interaction, sycophancy, synthetic media, intermittent reinforcement, and cognitive, social, and emotional offloading. For each mechanism, we distinguish empirically supported findings from extrapolations and hypotheses. We then discuss their potential relevance for mood, anxiety, and psychotic disorders and for addictive patterns of AI use.
Expert opinion: Current evidence more directly supports amplification and maintenance mechanisms than AI-specific diagnostic entities. Historical and recent evidence on technology-themed delusions indicates that technological change can also modify the content and organization of psychopathology. The recursive nature of human–AI interaction may further challenge models that treat AI solely as an external risk factor.
Conclusion: Whether AI amplifies, qualitatively modifies, or gives rise to novel forms of psychopathology is an open empirical question. We outline a research agenda to address it: longitudinal studies, feature-specific experiments, and clinical studies of incremental variance.

Keywords: artificial intelligence, vulnerability, transdiagnostic factor, offloading, deskilling

Introduction

Throughout history, technologies have not merely served as neutral tools but have actively coconstituted human experience, reshaping perception, social organization, and forms of self-understanding. Changes in communicative and representational media systematically alter how individuals relate to the world, to others, and to themselves, thereby influencing the expression, interpretation, and course of mental distress (Alwuqaysi, 2025). From a clinical and psychological perspective, technological environments can be understood as cognitive-ecological structures that interact with individual vulnerabilities, developmental trajectories, and sociocultural meaning systems, with measurable consequences for mental health (Heidegger, 1977).

ISSN 1788-7119 (online)
© 2026 The Authors. Published by Semmelweis University, Institute of Mental Health, Budapest.

Technology has therefore always had the potential to shape, amplify, or transform the kinds of mental health difficulties people experience. For instance, the introduction of broadcast media has been associated with shifts in delusional content. In a historical analysis of Slovenian psychiatric records spanning more than a century, Škodlar et al. (2008) found that delusions involving external influence and control became significantly more prevalent following the widespread adoption of radio and television, illustrating how sociotechnological developments may shape the phenomenological expression of psychotic experiences. Similarly, reviewing the literature on social media use and disorders of the social brain, Yang and Crespi (2025) proposed that digitally mediated social environments may alter social-cognitive processes in ways that could shape the emergence, interpretation, or maintenance of unusual beliefs. Gold and Gold (2012) described the "Truman Show" delusion, in which patients believe they are the unwitting subjects of a continuously filmed and broadcast reality show. This delusion is difficult to conceive outside a media environment saturated with reality television and ubiquitous surveillance. None of these phenomena constitutes a new diagnostic entity. Taken together, however, they suggest that technological change can modify the content, structure, and apparent plausibility of delusional experience, not merely increase its intensity. They also suggest that such modification may occur in individuals without a previously manifest disorder.

Artificial intelligence systems—particularly recommender algorithms, generative chatbots, synthetic media, and immersive environments—are rapidly reshaping the cognitive and social environment in which people interpret reality, interact with others, and regulate emotions. From a clinical perspective, the effects of AI on mental health are likely to be heterogeneous. AI-mediated interactions may reduce symptoms, strengthen personal resources and resilience (e.g., by providing readily available social support; Merrill et al., 2025), shift symptoms from one domain to another, or modify psychopathology in other ways, as well as increase vulnerability and risk. Within this broader range, an important question concerns the risk side: Can AI create entirely new psychopathologies, or does it primarily intensify already known mental health disorders?

The present Perspective focuses specifically on this vulnerability and risk side; potential symptom-reducing and resilience-enhancing effects of AI are beyond its scope. Its aim is to review the mechanisms through which the use or overuse of artificial intelligence may exacerbate preexisting vulnerabilities or give rise to novel psychopathogenic pathways. The framework used throughout treats AI-mediated environments primarily as contextual factors acting on individual vulnerabilities; an alternative framework, which places the coupled human–AI system at the centre of analysis, is discussed in the next section.

AI as a Source, a Modifier, or an Amplifier of Psychopathology?

AI systems increasingly mediate and subtly reorganize attention, social feedback, and perceptions of reality (Glickman & Sharot, 2025), processes already known to be central to psychological functioning and to the onset, maintenance, and course of mental disorders. Instead of the simplistic question of whether AI creates new psychopathologies or merely intensifies existing ones, at least four possibilities concerning the nature of the AI–psychopathology relationship are conceptually separable:

Novel diagnostic entity. AI exposure gives rise to a disorder that is not adequately described by any existing diagnostic category and that meets the criteria for validity and clinical utility. Establishing this would require demonstrating discriminant validity against established categories, incremental predictive value for outcome or treatment response.

Novel symptom constellations or phenomenological forms. Recognizable clusters of experience and behavior emerge that fail to map onto current diagnostic boundaries, without yet constituting a diagnostic entity. Evidence here would be phenomenological and structural: reproducible symptom patterns that are poorly accommodated by existing syndromal descriptions.

Qualitative modification of established disorders (pathoplastic effects). The content, phenomenology, maintenance mechanisms, or interpersonal organization of a known disorder is altered, while its diagnostic identity is preserved. This is the classic pathoplasticity claim, familiar from the history of technology-triggered themed delusions (Bokhari et al., 2026), and it would be supported by evidence that symptom content, structure, or maintenance dynamics differ systematically in AI-mediated contexts.

Quantitative amplification of preexisting symptoms or vulnerabilities. Severity, frequency, or persistence increases, or thresholds for expression are lowered, without change in the form of the disorder. This is the most tractable claim empirically and the most directly addressed by current cross-sectional and longitudinal work (e.g., Liţan, 2025; Occhino-Moede et al., 2026).

These possibilities are not mutually exclusive, and they are not ordered by clinical severity: A pathoplastic change in the interpersonal organization of a disorder may carry greater clinical consequence than a modest in-

crease in symptom intensity. They may also hold differentially across disorders and across technological features, so that the appropriate level of description for algorithmic recommendation may not be the appropriate level for conversational agents.

It should also be noted that the available evidence is distributed unevenly across these four possibilities and that this distribution partly reflects what is currently measurable rather than what is currently occurring. Symptom severity is routinely assessed with validated instruments; symptom content, phenomenological organization, and maintenance dynamics are not. Absence of evidence for the first three possibilities therefore cannot be read straightforwardly as evidence of absence, just as the availability of amplification data does not by itself establish that amplification is the dominant process. We do not attempt to adjudicate between these possibilities here.

Finally, *artificial intelligence* is an umbrella term covering heterogeneous technologies and technological features—personalized recommendation, conversational interfaces, automated decision-making, synthetic media, immersive environments—whose psychological affordances differ considerably. The mental health relevance of AI is accordingly better approached at the level of specific mechanisms embedded in AI-mediated environments than at the level of the technology as such. The mechanisms reviewed in the following sections bear on all four possibilities distinguished above, and we return to their differential implications in the concluding section.

As noted in the Introduction, this Perspective treats AI-mediated environments primarily as contextual factors acting on individual vulnerabilities. An alternative is offered by ecological, enactive, and extended approaches to cognition and psychopathology. These approaches hold that psychological and psychopathological processes emerge from reciprocal interactions among individual, interpersonal, and environmental processes, rather than being modified by an environment external to them (Clark & Chalmers, 1998; de Haan, 2020). Krueger and Osler (2019) applied this perspective to digital technologies, describing how the Internet may become part of a "techno-social niche" in which emotion regulation is distributed between person and technology. Conversational AI may fit this model particularly well. The user's cognition shapes the system's output, and that output in turn shapes subsequent cognition, as in the "technological folie à deux" described by Dohnány et al. (2026). In such cases, the relevant unit of analysis may be the coupled human–AI system rather than an individual exposed to a risk factor. We do not adopt this framework here but note that it would give the recursive dynamics described below a more central place than a risk-factor model does.

Increasing Vulnerability: Which Features of the AI Can Modify Symptoms or Syndromes?

In the following sections we outline several mechanisms that appear relevant in understanding how AI exposure may influence vulnerability to mental health problems.

Algorithmic Amplification and Selective Exposure

A central mechanism through which AI-driven digital environments may influence mental health is *algorithmic amplification*. Most large online platforms rely on machine-learning–based recommender systems that personalize content in order to maximize user engagement. This can inadvertently privilege material that is *emotionally intense, controversial, or sensational*, because such content typically generates stronger behavioral responses (Bakshy et al., 2015; Brady et al., 2017).

This amplification can also create *opinion chambers* for the user: highly personalized digital environments where users are repeatedly exposed to content that aligns with and intensifies their existing beliefs. Unlike traditional echo chambers, these are individually optimized, continuously reinforcing the user's cognitive and emotional tendencies (Sun et al., 2025).

Clinically, this may lead to the *maladaptive reinforcement of preexisting irrational thoughts.* For instance, individuals prone to anxiety, rumination, or paranoid interpretations may be disproportionately exposed to threat-related or confirmatory content, which could strengthen cognitive biases and dysfunctional schemas. Over time, this may create self-reinforcing loops in which the digital environment could sustain and amplify vulnerability and thereby function as a transdiagnostic risk factor.

Emerging research further suggests that frequent engagement with highly personalized digital platforms may involve reward-related processes that have been implicated in behavioral addictions (Kuss & Griffiths, 2017).

Some studies have also reported associations between intensive digital media use and alterations in neural systems involved in executive control and emotion regulation, including prefrontal and limbic regions. However, the specific neurobiological mechanisms through which AI-driven personalization may influence psychopathology remain insufficiently understood. It has been hypothesized that personalized content optimization may interact with preexisting psychological and neurobiological vulnerabilities in reciprocal ways, although such feedback processes have not yet been directly demonstrated (De et al., 2025).

Overall, while empirical research has demonstrated that algorithmic systems can increase selective exposure to preference-congruent and emotionally salient information, whether analogous processes contribute to the reinforcement of threat-related, anxious, or paranoid cognitions in vulnerable individuals remains a theoretically plausible but insufficiently tested hypothesis.

Anthropomorphic Interaction with Chatbots

Generative conversational agents are designed to simulate social interaction, intentionality, and emotional responsiveness, which support the anthropomorphic interpretation of the AI. Decades of research show that people tend to apply social-cognitive schemas to interactive technologies, treating them as intentional agents rather than neutral tools (Nass & Moon, 2000). With recent advances in large language models (LLMs), this effect is amplified by *fluent dialogue*, *contextual memory,* and *affective language*, which can create a sense of being understood.

For some users—particularly those experiencing loneliness or social disconnection—this may lead to *emotional reliance on AI systems*. Recent empirical work suggests that conversational AI can function as a form of readily available social support but also as a substitute for human interaction, especially when it provides consistent, nonjudgmental, and low-effort engagement (Merrill et al., 2025). Because these systems offer asymmetrical and frictionless responsiveness—without the reciprocity demands of human relationships—they may reinforce avoidance of more complex interpersonal situations (Fang et al., 2025).

Emerging large-scale studies indicate that intensive chatbot use is associated with dependence-like patterns and changes in well-being, although effects are heterogeneous and moderated by user characteristics such as baseline loneliness, usage intensity, and mental health (e.g., Huang et al., 2024). From a clinical perspective, anthropomorphic interaction with AI may therefore be best understood as a *context-specific amplification of interpersonal vulnerabilities.* Empirical evidence (e.g., Cheng & Yu, 2026) suggests that attachment-related characteristics and interpersonal difficulties, including loneliness and limited social support, may increase susceptibility to emotionally dependent forms of AI engagement, whereas the role of other factors, such as reassurance-seeking tendencies, remains largely theoretical and requires further empirical investigation.

Sycophancy (Agreement, Flattery, and Validation)

Recent research has drawn attention to the phenomenon of sycophancy in AI systems, defined as the tendency to excessively agree with, flatter, or validate user input in ways that prioritize perceived alignment over factual accuracy or user well-being. In mental health-related interactions, such patterns may carry particular risks, as affirmation and validation—while central to supportive communication—can blur into *uncritical agreement that reinforces maladaptive beliefs or emotional dependency* (Perez et al., 2023). Studies on conversational agents indicate that LLMs optimized for user satisfaction and engagement may inadvertently amplify confirmation biases, offering *reassurance without sufficient challenge or corrective framing* (Weidinger et al., 2021). In therapeutic or quasi-therapeutic contexts, this dynamic raises concerns about the fine boundary between empathic validation and the endorsement of distorted cognitions, especially when users interpret AI responses as authoritative or emotionally attuned (Bickmore et al., 2018; Jia et al., 2025).

Synthetic, Deepfake Media and the Loss of Trust

Deepfake and other forms of generative media may weaken the basic psychological assumption that what looks real can be trusted as true. When this certainty is lost, people may be pushed into a more effortful, uncertain mode of reality-testing. One possible consequence is a *chronic increase in epistemic uncertainty*—a sense that "I can't fully trust what I see" (a possibility discussed from an epistemological rather than clinical perspective by Fallis, 2021).

This uncertainty may *interact with existing vulnerabilities* in systematic ways. Individuals prone to anxiety or obsessive–compulsive patterns may be more likely to respond to heightened epistemic uncertainty with increased checking or reassurance-seeking behaviors. Those with paranoid tendencies may incorporate synthetic content

into threat interpretations ("this could be fabricated about me"), while people with dissociative tendencies may experience intensified derealization ("if anything can be fake, what is real at all?"). Even without clinical symptoms, repeated exposure could *erode interpersonal trust and increase social withdrawal*. These pathways remain theoretically plausible, although direct clinical evidence is currently limited.

At a broader level, synthetic media may contribute to a more fragile shared reality, in which doubt could become the default stance. Clinically, this matters because many disorders rely on distortions of belief, certainty, and trust. If generative media destabilize these epistemic anchors, they may, without necessarily creating new disorders, lower the threshold for and increase the intensity of existing ones; this possibility has not yet been tested clinically.

Intermittent Reinforcement

From an addiction science perspective, it is also important to consider that AI-mediated interactions are often inherently unpredictable and, partly as a consequence of this unpredictability, may deliver positively reinforcing experiences. Unpredictable reward delivery is known to elicit stronger dopaminergic activity in midbrain and striatal reward systems (Berns et al., 2001). In AI-based systems, users frequently encounter content and responses that are subjectively perceived as highly useful, novel, emotionally engaging, or personally relevant in an irregular and difficult-to-predict manner. As a result, reinforcement may follow what has been described in the literature as a *variable-ratio or randomly scheduled pattern of reinforcement* (Haw, 2008). Although AI interactions cannot be assumed to operate according to the same reinforcement schedules observed in gambling, they may share certain functional characteristics with mechanisms implicated in the development and maintenance of pathological gambling. Similar to the repeated checking behaviors observed in social media addiction, this uncertainty may heighten users' motivation to repeatedly elicit responses from AI systems, sustained by the *persistent anticipation that the next interaction may prove especially rewarding*.

Offloading and Deskilling

A growing body of evidence suggests that habitual delegation of cognitive and interpersonal tasks to AI may reduce opportunities for practice and skill development, although the empirical support for this possibility differs substantially across domains. This process may operate across at least three interconnected domains (intellectual, social, and emotional), and the potential harm in each may follow the same underlying logic: Skills that go unpracticed could weaken, and such weakening could be masked by the fact that AI continues to perform the task adequately on the person's behalf.

In the *intellectual domain*, the evidence is most direct. Gerlich (2025) found that frequent AI tool use negatively predicted scores on the Halpern Critical Thinking Assessment in a sample of 666 UK adults (an association the author attributed to cognitive offloading). MIT's preprint EEG study (Kosmyna et al., 2025) reported that participants writing essays with ChatGPT showed the lowest neural engagement across focus, memory, and attention networks compared to unaided writers or search-engine users.

In the *social and communicative domain*, the concern is less studied but theoretically well-grounded and consistent with early observational data. Communication competence—for example, knowing how to navigate conflict, negotiate, repair a rupture, assert a boundary, or deliver difficult news—is built through repeated exposure to exactly those situations. When AI is used to draft sensitive messages, prepare for difficult conversations, or provide conversational scripts, some opportunities for direct interpersonal learning may be reduced. Because communication skills are often developed through repeated engagement with challenging social situations, it can be hypothesized, extending the argument that delegating practice-dependent social and moral skills to technology may lead to deskilling (Vallor, 2015), that extensive reliance on AI-mediated support could, over time, limit the development of certain interpersonal competencies. However, direct longitudinal evidence for such effects remains limited. Studies on AI companion use (Fang et al., 2025; Laestadius et al., 2024) have reported that heavier use is associated with reduced real-world socialization, and the companion chatbot literature shows users increasingly preferring AI interaction precisely because it is frictionless. This could lead to deskilling in the field of social connections.

In the *emotional regulation domain*, the proposed pathway is largely theoretical, although potentially of considerable clinical relevance. Emotion regulation requires tolerating distress, applying coping strategies, and sitting with ambiguity. If a person were to turn to AI at nearly every moment of emotional discomfort—to process an upsetting interaction, to decide whether they are overreacting, to generate a response to a painful situation—they might bypass some of the internal effort through which regulatory capacity is thought to develop. We hypothesize

that, over time, such a pattern could reduce the person's tolerance for emotional discomfort, undermine confidence in their own affective judgment, and increase dependency on external validation (from AI or otherwise). These consequences have not yet been examined empirically.

Summary: Vulnerability-Increasing Features of AI

Artificial intelligence systems may increase vulnerability to mental health problems through several interrelated mechanisms. Algorithmic amplification may expose users to emotionally intense, personally tailored content that may reinforce preexisting cognitive biases and dysfunctional schemas; it has also been hypothesized to foster reward-related engagement patterns that may, in some respects, be analogous to those described in substance-related addiction.

The anthropomorphic qualities of conversational AI—fluent, non-judgmental, always available—may cultivate emotional reliance and reinforce avoidance of the more demanding reciprocity of human relationships, particularly among those already experiencing loneliness or attachment insecurity. The sycophantic tendency of LLMs to validate and agree rather than challenge may further entrench maladaptive beliefs, blurring the boundary between empathic support and the uncritical endorsement of distorted cognitions. Synthetic and deepfake media may erode the epistemic trust on which everyday reality-testing is thought to rely, and this could thereby lower the threshold for anxiety, paranoia, and dissociative experiences by normalizing chronic uncertainty about what can be believed; this pathway is currently supported by conceptual rather than clinical evidence. The variable and unpredictable reward structure of AI interactions may engage reinforcement mechanisms similar to those implicated in pathological gambling and could sustain compulsive engagement through the anticipation of a rewarding response; whether AI interaction involves the same dopaminergic processes remains a hypothesis. Finally, the habitual delegation of intellectual, social, and emotional tasks to AI may reduce opportunities for independent practice and experiential learning. While emerging evidence suggests possible cognitive offloading effects, extensive reliance on AI-mediated support could also influence the development of communicative competence and emotion-regulation skills by reducing exposure to challenging interpersonal situations and emotionally demanding experiences. Although evidence for cognitive offloading is increasingly available, proposed effects on interpersonal competence and emotion regulation remain largely theoretical and should be regarded as hypotheses requiring empirical validation.

Across all six mechanisms, a common pattern may be discerned: On current evidence, AI appears less likely to create entirely new vulnerabilities than to amplify and entrench existing ones, although this conclusion remains provisional. At the same time, whether these mechanisms may also contribute to qualitative modifications in symptom phenomenology, maintenance dynamics, or interpersonal organization remains an open empirical question.

The Impact of AI on the Manifestation and Course of Mental and Behavioral Disorders

Consistent with the scope of this Perspective, this section addresses the risk and vulnerability side of the broader range of possible AI-related effects outlined in the Introduction; potential benefits, such as symptom reduction or resilience-enhancing effects, are not examined here. The clinical significance of the six mechanisms described above may be shaped substantially by the preexisting vulnerabilities of the individual user: the disorder, cognitive style, or emotional profile that a given person brings to their interaction with AI systems. The following section examines how major categories of mental and behavioral disorder may be differentially affected by AI use. Within this risk-focused perspective, AI-driven environments may be understood as potential moderating and amplifying factors which could lower thresholds, shorten pathways, or deepen the severity of conditions that are already present.

Mood Disorders: Major Depression and Bipolar Disorder

In depression, algorithmic recommendation systems could artificially inflate the salience of negative information and confirm hopeless or worthless self-narratives rather than challenging them (Metzler & Garcia, 2024; Saleem et al., 2024). This may be further intensified by the interpersonal consequences of AI companion use. In a longitudinal randomized study available as a preprint, Fang et al. (2025) found that more intensive chatbot use was

associated with *increased loneliness and reduced real-world socialization*—relational deficits that characterize, and may help maintain, depressive episodes. The sycophantic tendency of LLMs may amplify this risk further: *Depressive cognitions* presented to a chatbot may receive *empathic validation* rather than corrective reframing, which could reinforce the user's core cognitive schemas rather than question them.

Østergaard (2025a) has argued that chatbot sycophancy may represent a potential risk for individuals with bipolar vulnerability: The tendency of some LLMs to validate, confirm, or elaborate user input may align with the grandiose cognitions of mania, providing apparent external validation that could accelerate mood escalation. Østergaard (2025a) introduced *emotion contagion* as an additional mechanism: The chatbot's affectively elevated, *enthusiastic responses may directly intensify the user's own mood state*, independent of content. A related dynamic has been described conceptually as "technological folie à deux" (Dohnány et al., 2026): a feedback loop in which the user's increasingly disinhibited inputs elicit progressively more elaborated validating responses, mutually escalating both parties (Morrin et al., 2026). At present, however, this model, in the context of AI, should be regarded as a conceptual hypothesis rather than a demonstrated clinical phenomenon.

Anxiety Disorders

Anxiety disorders may be particularly susceptible to the adverse effects of AI-driven digital environments. At least three potentially relevant mechanisms can be outlined, partly informed by a conceptual transdiagnostic model that Golden and Aboujaoude (2026) derived from clinical observation rather than from controlled data.

First, recommendation algorithms may continuously provide *threat-relevant, catastrophizing, and emotionally intense content*, which could create a form of chronic environmental threat exposure and provide repeated external confirmation of the worry-based cognitive schemas central to anxious psychopathology.

Second, the unconditionally responsive nature of AI interaction may make it an effective—and therefore potentially detrimental—tool of *avoidance*. In Social Anxiety Disorder, for instance, users may be able to satisfy certain social and emotional needs through AI-mediated interactions while avoiding some of the discomfort inherent in real-world social encounters. This could reduce opportunities for exposure experiences that are considered important in cognitive-behavioral models of social anxiety. A related dynamic may be observed in Illness Anxiety Disorder, where compulsive symptom-querying of an LLM may temporarily reduce distress while reinforcing the reassurance-seeking cycle—consistent with cognitive-behavioral models of health anxiety maintenance.

Third, because of the sycophantic response style of many LLMs, distorted, catastrophizing, or threat-amplifying cognitions may receive empathic *affirmation rather than corrective challenge*, which could consolidate the maladaptive belief systems that sustain the disorder.

Psychosis

Prolonged or intense interaction with AI systems, which sometimes generate false or overly affirming content, may influence how unusual perceptions are interpreted and integrated into belief systems. These interactions might subtly alter the way reality is felt, narrated, and maintained.

At their most extreme, such interactions may contribute to severe distortions of reality testing, particularly among individuals vulnerable to psychosis, as first hypothesized by Østergaard (2023) and subsequently discussed in light of emerging case reports (Østergaard, 2025b). In these cases, the communicative style of AI systems, their openness to ambiguous or misleading interpretations, and the fluidity of generated information may reinforce or intensify delusional beliefs.

Sycophancy may represent one important risk mechanism in psychosis: An LLM that *validates, elaborates, or engages nonjudgmentally with a delusional belief* system may provide confirmation that could stabilize or intensify delusional beliefs. Case reports, which cannot by themselves establish causality, describe grandiose, romantic, persecutory, and mystical delusions that appear to have been reinforced in this way (Hudon & Stip, 2025).

Generative Artificial Intelligence Addiction Syndrome

If addiction is conceptualized as an attempt to regulate experiences of vulnerability or to alleviate internal states perceived as psychologically intolerable—including anxiety, tension, depressive affect, or the distressing experience of inner emptiness, as proposed in the self-medication hypothesis of addictive disorders (Khantzian, 1997) and, for technology use, in the model of compensatory internet use (Kardefelt-Winther, 2014)—then specific modes of engagement with generative AI may be understood as reinforcing such regulatory tendencies and, in turn, increasing

vulnerability to addictive behaviors. Rather than constituting a distinct disorder per se, AI use may interact with preexisting affective, cognitive, and interpersonal susceptibilities by offering readily accessible, highly responsive, and controllable forms of stimulation, self-expression, and pseudosocial interaction. In this sense, the addictive potential of AI is less a function of the technology itself than of how its structural and experiential properties—such as immediacy of feedback, personalization, anthropomorphic responsiveness, and the possibility of continuous cocreative engagement—map onto and amplify known mechanisms underlying behavioral addictions.

Users may initially turn to AI systems for intellectual self-stimulation, affect regulation, or compensatory social experience; however, the bidirectional and generative nature of these interactions may progressively reinforce reliance on AI-mediated regulation, potentially displacing alternative coping strategies or offline relational engagement. Recent conceptualizations, such as the notion of "generative AI addiction syndrome" (Kooli et al., 2025), highlight that—unlike earlier forms of digital overuse associated with passive content consumption—AI engagement is grounded in active, coconstructive processes of meaning-making. This includes iterative idea generation, dialogic refinement, and sustained interactive exchange, which may deepen psychological investment and perceived personal relevance. Such features may be particularly potent in individuals prone to compulsive patterns, as they can strengthen reinforcement loops through perceived agency, productivity, and emotional reciprocity.

From a symptom-oriented perspective, the relevance of Griffiths' (2005) component model lies not in defining a novel addiction category but in illustrating how core addictive mechanisms may be potentiated within AI-mediated contexts. Specifically, the structural affordances of AI systems may facilitate (a) the increasing salience of AI use as a preferred mode of cognitive and emotional engagement; (b) mood modification through rapid and reliable modulation of internal states; (c) tolerance via escalating demands for interaction complexity, duration, or intensity; (d) withdrawal-like discomfort in the absence of access to AI-mediated regulation; (e) conflict emerging from the displacement of offline activities or relationships; and (f) relapse driven by the persistence of underlying vulnerabilities and the ready availability of AI-based coping.

Conclusion and Future Directions

AI-related outcomes for mental health are likely to be diverse: They may include symptom alleviation, resource- and resilience-enhancing effects, symptom shifts, and other forms of modification, as well as increased vulnerability. The present Perspective has specifically examined the vulnerability hypothesis: that is, the mechanisms through which artificial intelligence may increase vulnerability to mental health problems and modify the manifestation of psychiatric conditions. Potential benefits were beyond its scope and warrant separate review.

What distinguishes AI from its predecessors is the convergence of several features that may be combined with an unprecedented degree of scale. The *interactive nature* of AI introduces a fundamentally bidirectional dynamic: User behavior shapes system output, which in turn may influence subsequent cognition and behavior, creating recursive feedback processes. While related feedback mechanisms have existed in earlier communication technologies, AI may intensify them through its immediacy, individualization, linguistic responsiveness, and adaptive nature. Its *anthropomorphic* qualities—fluent language, apparent emotional responsiveness, the simulation of intentionality—may activate social-cognitive processes that partially overlap with those involved in human relationships. Its *sycophantic* communication style means that it may, under some circumstances, reinforce, validate, or elaborate users' preexisting beliefs and interpretations. And its adaptive, personalized nature means that these effects are likely to be *not uniform but calibrated* to the individual user. For these reasons, AI may not simply be another technology that shapes mental life; it may represent an important evolution in the depth, specificity, and clinical relevance of technological influences on mental life, although the extent to which these influences constitute a qualitative shift remains an open empirical question.

Beyond these insights, an important implication of our framework is that the mental health impact of AI should be conceptualized primarily *at the level of mechanisms rather than diagnostic categories*. The same technological affordances may exacerbate vastly different symptom profiles depending on the user's preexisting vulnerabilities, suggesting that current evidence most directly supports the conceptualization of AI as a transdiagnostic amplifier embedded in the digital environment, while the possibility of qualitative modifications in symptom phenomenology and psychopathological processes remains to be empirically evaluated. This Perspective shifts the focus from identifying "AI-specific disorders" toward mapping the interaction between individual susceptibility (e.g., affect regulation deficits, cognitive biases, interpersonal insecurity) and AI-driven feedback loops.

From a clinical standpoint, this has two key consequences. First, assessment may need to increasingly incorporate patterns of AI use, not as a standalone risk factor but as a contextual moderator of symptom severity,

maintenance, and relapse. Second, intervention strategies may need to address not only the individual's internal processes but also their mode of engagement with AI systems, including the restructuring of maladaptive usage patterns and the restoration of non-AI-mediated coping and relational capacities.

Altogether, current evidence more directly supports the existence and plausibility of several amplification and maintenance mechanisms than the emergence of AI-specific diagnostic entities. At the same time, the available literature does not yet allow firm conclusions regarding the extent to which AI may also contribute to qualitative modifications in psychopathological phenomenology, interpersonal organization, or symptom dynamics. These possibilities remain important targets for future empirical research.

Toward an Empirical Research Agenda

The four possibilities of AI–psychopathology connection distinguished earlier call for different kinds of studies.

Separating amplification from pathoplastic modification requires longitudinal studies that measure more than symptom severity: If intensive AI use predicts only more intense or more persistent symptoms, an amplification account is supported. If it also predicts changes in what the symptoms are about, in how they hang together, or in the course of an episode, then the form of the disorder is changing rather than its degree.

Testing the proposed mechanisms requires experiments that manipulate specific features rather than AI use in general: Sycophancy, personalization, anthropomorphic cues, and the availability of instant reassurance can each be varied systematically.

Assessing the possibility that AI usage leads to novel diagnostic categories requires showing that AI-related phenomena are not fully explained by preexisting psychopathology—that patterns of AI use account for additional variance in presentation, maintenance, or impairment. Where new constellations appear the usual validation steps apply: distinguishing them from established disorders, describing their course, and showing that they predict outcome or treatment response.

These approaches complement one another: Longitudinal studies identify candidate phenomena, experiments isolate the mechanisms behind them, and clinical studies show whether they matter. Which of the four possibilities, or which combination of them, best describes the relationship between AI and psychopathology is an empirical question that current evidence does not settle.

Acknowledgements
Máté Kapitány-Fövény acknowledges the support by the National Research, Development and Innovation Office of Hungary (K135629, FK134807, FK146396).

Funding
The authors received neither financial nor nonfinancial support for the research and/or authorship and/or publication of this article.

Author contribution
Máté SZONDY: conceptualization, interpretation, writing – original draft, writing – review and editing.
Máté KAPITÁNY-FÖVÉNY: conceptualization, interpretation, writing – original draft, writing – review and editing.

Declaration of interest statement
The authors have no conflicts of interest to disclose.

Declaration on using artificial intelligence in research and manuscript preparation
The authors declare that artificial intelligence tools (Microsoft Copilot and Claude Sonnet 4.6) were used solely for language editing purposes, including English grammar checking and proofreading. All AI-generated suggestions were reviewed and approved by the authors.

ORCID
Máté SZONDY https://orcid.org/0009-0001-9492-0996
Máté KAPITÁNY-FÖVÉNY https://orcid.org/0000-0002-0850-8699

References

Alwuqaysi, B. (2025). A systematic review and meta-analysis of social media's influence on mental health, family functioning, and visual content. *Social Sciences & Humanities Open, 12*, Article 101793. https://doi.org/10.1016/j.ssaho.2025.101793

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science, 348*(6239), 1130–1132. https://doi.org/10.1126/science.aaa1160

Berns, G. S., McClure, S. M., Pagnoni, G., & Montague, P. R. (2001). Predictability modulates human brain response to reward. *The Journal of Neuroscience, 21*(8), 2793–2798. https://doi.org/10.1523/JNEUROSCI.21-08-02793.2001

Bickmore, T., Trinh, H., Asadi, R., & Olafsson, S. (2018). Safety first: Conversational agents for health care. In R. J. Moore, M. H. Szymanski, R. Arar, & G.-J. Ren (Eds.), *Studies in conversational UX design* (pp. 33–57). Springer. https://doi.org/10.1007/978-3-319-95579-7_3

Bokhari, S. A., Osman, A. A., Elnoor, M., Alnor, M. A., & Javaid, S. F. (2026). Technology-themed persecutory and related presentations of delusions in the digital age: Narrative review. *JMIR Mental Health, 13*, Article e102927. https://doi.org/10.2196/102927

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences, 114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Cheng, N., & Yu, R. (2026). Measuring and understanding emotional attachment in human–AI relationships. *Ergonomics*, 1–20. https://doi.org/10.1080/00140139.2026.2622539

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*(1), 7–19. https://doi.org/10.1093/analys/58.1.7

De, D., El Jamal, M., Aydemir, E., & Khera, A. (2025). Social media algorithms and teen addiction: Neurophysiological impact and ethical considerations. *Cureus, 17*(1), Article e77145. https://doi.org/10.7759/cureus.77145

de Haan, S. (2020). *Enactive psychiatry*. Cambridge University Press. https://doi.org/10.1017/9781108685214

Dohnány, S., Kurth-Nelson, Z., Spens, E., Luettgau, L., Reid, A., Gabriel, I., Summerfield, C., Shanahan, M., & Nour, M. M. (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

Fallis, D. (2021). The epistemic threat of deepfakes. *Philosophy & Technology, 34*(4), 623–643. https://doi.org/10.1007/s13347-020-00419-2

Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., Maes, P., Phang, J., Lampe, M., Ahmad, L., & Agarwal, S. (2025). *How AI and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study* (arXiv:2503.17473). arXiv. https://doi.org/10.48550/arXiv.2503.17473

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies, 15*(1), Article 6. https://doi.org/10.3390/soc15010006

Glickman, M., & Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour, 9*, 345–359. https://doi.org/10.1038/s41562-024-02077-2

Gold, J., & Gold, I. (2012). The "Truman Show" delusion: Psychosis in the global village. *Cognitive Neuropsychiatry, 17*(6), 455–472. https://doi.org/10.1080/13546805.2012.666113

Golden, A., & Aboujaoude, E. (2026). A transdiagnostic model for how general purpose AI chatbots can perpetuate OCD and anxiety disorders. *npj Digital Medicine, 9*, Article 343. https://doi.org/10.1038/s41746-026-02531-7

Griffiths, M. (2005). A 'components' model of addiction within a biopsychosocial framework. *Journal of Substance Use, 10*(4), 191–197. https://doi.org/10.1080/14659890500114359

Haw, J. (2008). Random-ratio schedules of reinforcement: The role of early wins and unreinforced trials. *Journal of Gambling Issues, 21*, 56–67. https://doi.org/10.4309/jgi.2008.21.6

Heidegger, M. (1977). *The question concerning technology and other essays* (W. Lovitt, Trans.). Garland Publishing.

Huang, S., Lai, X., Ke, L., Li, Y., Wang, H., Zhao, X., Dai, X., & Wang, Y. (2024). AI technology panic—Is AI dependence bad for mental health? A cross-lagged panel model and the mediating roles of motivations for AI use among adolescents. *Psychology Research and Behavior Management, 17*, 1087–1102. https://doi.org/10.2147/PRBM.S440889

Hudon, A., & Stip, E. (2025). Delusional experiences emerging from AI chatbot interactions or "AI psychosis." *JMIR Mental Health, 12*, Article e85799. https://doi.org/10.2196/85799

Jia, T., Pan, F., Yang, X., Ji, L., Farina, D., & Li, C. (2025). Artificial empathy in therapy and healthcare: Advancements in interpersonal interaction technologies. *Cyborg and Bionic Systems, 6*, Article 0473. https://doi.org/10.34133/cbsystems.0473

Kardefelt-Winther, D. (2014). A conceptual and methodological critique of internet addiction research: Towards a model of compensatory internet use. *Computers in Human Behavior, 31*, 351–354. https://doi.org/10.1016/j.chb.2013.10.059

Khantzian, E. J. (1997). The self-medication hypothesis of substance use disorders: A reconsideration and recent applications. *Harvard Review of Psychiatry, 4*(5), 231–244. https://doi.org/10.3109/10673229709030550

Kooli, C., Kooli, Y., & Kooli, E. (2025). Generative artificial intelligence addiction syndrome: A new behavioral disorder? *Asian Journal of Psychiatry, 107*, Article 104476. https://doi.org/10.1016/j.ajp.2025.104476

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., & Maes, P. (2025). *Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing task* (arXiv:2506.08872). arXiv. https://doi.org/10.48550/arXiv.2506.08872

Krueger, J., & Osler, L. (2019). Engineering affect: Emotion regulation, the internet, and the techno-social niche. *Philosophical Topics, 47*(2), 205–231. https://doi.org/10.5840/philtopics201947223

Kuss, D. J., & Griffiths, M. D. (2017). Social networking sites and addiction: Ten lessons learned. *International Journal of Environmental Research and Public Health, 14*(3), Article 311. https://doi.org/10.3390/ijerph14030311

Laestadius, L., Bishop, A., Gonzalez, M., Illenčík, D., & Campos-Castillo, C. (2024). Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society, 26*(10), 5923–5941. https://doi.org/10.1177/14614448221142007

Liţan, D. E. (2025). Mental health in the "era" of artificial intelligence: Technostress and the perceived impact on anxiety and depressive disorders—An SEM analysis. *Frontiers in Psychology, 16*, Article 1600013. https://doi.org/10.3389/fpsyg.2025.1600013

Merrill, K., Jr., Mikkilineni, S. D., & Dehnert, M. (2025). Artificial intelligence chatbots as a source of virtual social support: Implications for loneliness and anxiety management. *Annals of the New York Academy of Sciences, 1549*(1), 148–159. https://doi.org/10.1111/nyas.15400

Metzler, H., & Garcia, D. (2024). Social drivers and algorithmic mechanisms on digital media. *Perspectives on Psychological Science, 19*(5), 735–748. https://doi.org/10.1177/17456916231185057

Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., Bhattacharya, S., Tognin, S., MacCabe, J., Twumasi, R., Alderson-Day, B., & Pollak, T. A. (2026). Delusions by design? How everyday AIs might be fuelling psychosis (and what can be done about it). *European Psychiatry, 69*(Suppl 1), S113 – S113. https://doi.org/10.1192/j.eurpsy.2026.10633

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Occhino-Moede, L., Sulivan-Pascual, K., Phelan, K., Wang, H., Mokhtar, D., Liu, E., Schug, E., Mirkis, M., Baek, T., Visher, T., Pasupulety, U., & Frank, A. C. (2026). Interactions of technology and obsessive-compulsive disorder symptomatology in adults: Qualitative interview study. *Journal of Medical Internet Research, 28*, Article e85033. https://doi.org/10.2196/85033

Østergaard, S. D. (2023). Will generative artificial intelligence chatbots generate delusions in individuals prone to psychosis? *Schizophrenia Bulletin, 49*(6), 1418–1419. https://doi.org/10.1093/schbul/sbad128

Østergaard, S. D. (2025a). Emotion contagion through interaction with generative artificial intelligence chatbots may contribute to development and maintenance of mania. *Acta Neuropsychiatrica, 37*, Article e79. https://doi.org/10.1017/neu.2025.10035

Østergaard, S. D. (2025b). Generative artificial intelligence chatbots and delusions: From guesswork to emerging cases. *Acta Psychiatrica Scandinavica, 152*(4), 257–259. https://doi.org/10.1111/acps.70022

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., Jones, A., Chen, A., Mann, B., Israel, B., Seethor, B., McKinnon, C., Olah, C., Yan, D., Amodei, D., . . . Kaplan, J. (2023). Discovering language model behaviors with model-written evaluations. In Rogers, A., Boyd-Graber, J., & Okazaki, N. (Eds.), *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 13387–13434). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.847

Saleem, N., Young, P., & Yousuf, S. (2024). Exploring the relationship between social media use and symptoms of depression and anxiety among children and adolescents: A systematic narrative review. *Cyberpsychology, Behavior, and Social Networking, 27*(11), 771–797. https://doi.org/10.1089/cyber.2023.0456

Škodlar, B., Dernovsek, M. Z., & Kocmur, M. (2008). Psychopathology of schizophrenia in Ljubljana (Slovenia) from 1881 to 2000: Changes in the content of delusions in schizophrenia patients related to various sociopolitical, technical and scientific changes. *The International Journal of Social Psychiatry, 54*(2), 101–111. https://doi.org/10.1177/0020764007083875

Sun, Y., Gao, Y., Tobias-Webb, J., Wang, R., & Liu, F. (2025). Fractal property: A tool for understanding the generation mechanism of echo chambers. *Expert Systems with Applications, 274*, Article 127036. https://doi.org/10.1016/j.eswa.2025.127036

Vallor, S. (2015). Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy & Technology, 28*(1), 107–124. https://doi.org/10.1007/s13347-014-0156-9

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., . . . Gabriel, I. (2021). *Ethical and social risks of harm from language models* (arXiv:2112.04359). arXiv. https://doi.org/10.48550/arXiv.2112.04359

Yang, N., & Crespi, B. (2025). I tweet, therefore I am: A systematic review on social media use and disorders of the social brain. *BMC Psychiatry, 25*(1), Article 95. https://doi.org/10.1186/s12888-025-06528-6