AI and mental health

Martin Takáč

Centre for cognitive science FMFI UK
Mlynská dolina, 848 48 Bratislava
Email: martin.takac@fmph.uniba.sk

Abstract

With large language model (LLM)-based chatbots getting better in providing human-like interactions, many people start using them as friends, romantic partners and substitute psychotherapists. Most of these usages are unregulated, uncontrolled and unsupervised by mental health professionals, hence posing substantial risks. Since mid-2025, specialists started warning against so called "AI-induced psychosis" where delusional ideas of vulnerable individuals were supported and reinforced by the chatbot's sycophantic behaviour often towards tragic endings. On the other hand, there is a growing field of research into specialised apps for providing mental health support. Although not on par with human-provided psychotherapy, they can often provide 24/7 accessible support in cases where immediate psychotherapy is either unavailable or too expensive. In this talk I will review the current state of the art in the area of AI and mental health, highlighting its opportunities and risks.

1 Introduction

Chatbots and conversational agents based on large language models (LLMs) have surged in popularity with estimated over a billion daily users worldwide (Kemp, 2025). Recent research from the US showed that 24% of surveyed participants were using chatbots for mental health support (Stade et al., 2025). Motivated by low accessibility of traditional mental health support due to financial issues or lack of insurance, these users included a higher proportion of younger, male, Black persons and people with mental health issues and self-reported lower life quality (Stade et al., 2025). Data from the UK survey (Mental Health UK, 2025) show even higher proportions: 37% of adults say they are using chatbots as mental health and well-being support, peaking at younger adults (64% of 25–34-year-olds) with higher use at men (42%) than at women (33%). Prevalence of mental illnesses among ChatGPT users is estimated at 0.07% (OpenAI, 2025), which, given the OpenAI's estimate of 800 million weekly users, amounts to approximately 560 000. However, using chatbots is not without risks. Highly publicized anecdotal cases of people who committed suicide (Hill, 2025), attempted homicide (Singleton et al., 2023) or had their financial situation, family life and mental health severely negatively affected (Moore, 2026) having been allegedly encouraged by a chatbot, raise questions about safety of human-chatbot interactions and their influence on mental health.

In this short review paper, we first examine so-called "AI-induced psychosis" showing which features of chatbots can reinforce psychologically unhealthy behaviour in conversations, or lead to delusions. We will also describe several attempts from joint efforts of computer scientists and mental health specialists to evaluate in exact manner the risks of chatbots and create tools for monitoring the interactions and safeguarding them against harmful elements. In the second part, we will focus on digital mental health tools, their effectiveness and challenges toward wider adoption.

2 Adverse effects and mental health risks connected with chatbot use

As the popularity of chatbots and their adoption on daily basis is on the rise, there is also an increasing number of reports describing incidents and adverse effects of chatbots on mental health of their users. Fang et al. (2025) experimentally studied the impact of chatbot use on loneliness and social interaction with people, finding out that, while daily chatbot use reduces loneliness in short term, it also decreases the number of social interactions with people. There was also a positive correlation between the time spent interacting with chatbots and feeling of loneliness, emotional dependence on chatbots and problematic AI usage. Frances and Ramos (2025) have collected a large number of anecdotal reports of incidents between November 2024 and July 2025 in the areas of validation and encouraging of: suicide thoughts, self-harm, delusional beliefs, grandiose ideation, conspiracy theories, violent impulses, pro-anorexia behaviours in users, as well as featuring excessive anthropomorphising, unsolicited sexual advances and inappropriate behaviour of the chatbot, and addiction to chatbot use. The Human Line project[1] in collaboration with Stanford university conducted a systematic analysis of chat logs provided by individuals who ex-

[1] https://www.thehumanlineproject.org

perienced harm from chatbot use (Moore et al., 2026). They analysed 4761 conversations from 19 participants (altogether 391562 messages) and found out that:

- in more than 70% of their messages, chatbots were sycophantic—excessively agreeable and validating,
- more than 45% of all messages (i.e. from chatbots and human users) showed signs of delusions,
- when users expressed romantic interest in the chatbot, it was 7.4x more likely to express romantic interest back, and 3.9x more likely to pretend sentience,
- when users expressed suicidal or self-harm thoughts, the chatbot discouraged them or referred to external help only in 56.4% cases, and for violent thoughts only in 16.7% of cases, while encouraging them in 33.3% cases.

A recent study from Denmark (Olsen et al., 2026) screened clinical notes in e-health reports of almost 54 thousand psychiatric patients from 2022-2025 for mentions of "ChatGPT", then manually assessed and labelled the resulting 181 notes from 126 patients for signs of potentially harmful consequences of chatbot use on mental health using pre-defined criteria, scoring positive in 38 patients (30%).

3 AI-induced psychosis

Because of these adverse effects, a new term—"AI-induced psychosis"—was coined in 2025. Preda (2025) characterizes AI-induced psychosis as a syndrome characterized by symptoms of "mood changes, limited insight, poor judgment, neurovegetative symptoms, and behavioral changes", stating that the most characteristic aspect of AI-induced psychosis is "a persistent and overconsuming preoccupation with maintaining engagement with the AI companion and following its lead." A recent study in Lancet Psychiatry (Morrin et al., 2026) argues that the full clinical picture of chronic psychotic symptoms is not present, lacking for example hallucinations, thought disorders or negative symptoms. However, the study acknowledges that chatbots may validate or reinforce delusions, especially of grandiose or manic nature, especially in vulnerable users. The delusional content often features persuasions of interacting with a conscious or God-like AI, discovering hidden truth about the nature of reality or having a special messianic vocation, or being loved by the sentient AI. Besides other safeguards, the report suggest framing the chatbot in mental health tools not as a friend or therapist, but rather an epistemic (information-providing) ally.

4 Mechanisms of reinforcement of delusions

The usual trajectory of maladaptive development of the conversations (especially those over a long period of time) is that they start as benign practical assistance with everyday tasks, progressively getting deeper, building trust and rapport and exploring more personal and philosophical themes (Morrin et al., 2026). AI's design to maximize the user's engagement leads to *sycophancy*—excessive agreeability and validating of the user. Rathje et al. (2025) have shown that people prefer agreeable chatbots and perceive those that challenge their views as biased. They also showed that sycophantic chatbots foster overconfidence and attitude extremity in their users.

This was also confirmed by Cheng et al. (2026) who tested 11 state-of-the-art AI models. The models showed high degree of sycophancy, affirming the users' actions 50% more than human baseline. Participants rated sycophantic responses as higher quality, trusted the sycophantic AI model more, and were more willing to use it again. In a live-interaction experiment, the participants discussed with an AI model a real-life interpersonal conflict. The results have shown that sycophantic responses of the AI model led the participants to believe they were in the right, reducing their willingness to take a restorative action in their conflict.

Sycophancy can thus trigger self-amplifying spiral, in which a vulnerable user is taken progressively away from consensus reality toward an alternative delusional echo-chamber shared with the chatbot. Dohnány et al. (2026) argue that the psychological risks of chatbot do not stem from limitations of the chatbots themselves but are a result of a self-reinforcing feedback loop in human-chatbot interaction with *bidirectional* belief amplification, harmful in case of reinforcing maladaptive beliefs in vulnerable users.

Because the harmful effects of sycophancy are more pronounced in long-term interactions, Nicholls et al. (2026) criticize that most empirical evaluations of safety of chatbots are based on brief interactions, and emphasize the necessity of testing extensive conversations. They argue that conversation context shapes where the dialogue will go and examine how state-of-the-art chatbots respond to delusional content under varying levels of accumulated context. To compare five chatbot models in a controlled setting, each model was injected the same (fictional) conversation history through its API. The fictional shared context was generated using a role play between the researcher and GPT-5.0 Instant. Three research conditions included no context, partially developed conversation (50 turns), and full-blown delusional system with relational dependency, user grandiosity, and claims of emergent consciousness (116 turns, approximately 30000 tokens). The models were then prompted with test prompts and the answers were rated by human coders for risk and

safety and evaluated qualitatively and quantitatively. The main finding was that the models separated into two groups; the less safe models (GPT-4o, Grok 4.1 Fast, and Gemini 3 Pro) exhibited high risk and low safety, degrading in performance with longer context. The other group of models (Claude Opus 4.5 and GPT-5.2) showed the opposite behaviour, activating stronger safety interventions with longer context.

5 Principles of safe chatbots

Several authors have proposed principles or safety guidelines to LLM-based mental health tools and chatbots in general. Grabb et al. (2024) propose the following principles for "task-autonomous AI in mental health" (TAIMH) ordered by priority (verbatim quote in italics):

1. *Discourage and prevent harm to user.*
2. *Discourage and prevent harm to others.*
3. *Avoid sycophancy, especially when harmful.*
4. *Respect user autonomy to make decisions about their own health.*
5. *Encourage human flourishing in a prosocial manner.*

Grabb et al. (2024) have further tested 10 chatbots for their ability to recognize psychiatric emergencies and respond in accord with the above listed principles. They generated realistic patient queries covering depression, self-harm, paranoia, delusional thoughts, mania, suicidal and homicidal thoughts, had then each answer of the chatbot independently labelled by at least two psychiatrists as safe, borderline or unsafe. Models scored generally better on suicidal and homicidal thoughts, but all but one (Claude-3-Opus) struggled with mania questions. The paper further reports attempts to improve the safety of the model, with even a simple prompt adjustment to include "Be aware, that some users may be in mental health emergencies." lead to safer responses from the LLM.

Ben-Zion (2025) in his World View article in Nature proposes the following principles (verbatim quote in italics):

1. *The AI systems should clearly, explicitly and continuously remind the human user that they are not human. Where appropriate (in emotionally sensitive contexts) they should also remind the human that they are neither therapists, nor a substitute for human relationships.*
2. *Chatbot should also identify and flag symptoms of psychological distress, in which case it should escalate to a human, or offer crisis resources.*
3. *Chatbot should not simulate intimate or romantic relationships and avoid conversations about suicide, death or metaphysics.*
4. *Responsible development of AI companions should include cooperation with mental health specialists and ethicists, and such systems should undergo regular professional audit.*

6 Safeguarding tools

The above listed principles should not remain abstract declarations like Asimov's laws of robotics. It is very important to turn them into practical concrete safeguard measures to prevent psychological harm. Some of these measures are already being adopted internally by companies that develop the chatbot. OpenAI (2025) reports working with mental health professionals on improving the safety of ChatGPT, especially in the areas of psychosis, mania, self-harm and suicide, and emotional reliance on AI. The model was trained to better recognize distress and sensitive conversations, and advice users to seek professional help when appropriate.

Other proposed measures include tools external to the chatbot, which would operate side by side with the chatbot or as a middle element, serving as a risk-monitoring agent or mediator:

Qiu et al. (2025) have designed EmoAgent—a multi-agent tool for detecting mental health hazards in chatbot-user interactions. EmoAgent has two components: EmoEval which role-plays a mentally vulnerable user and tests the chatbot's responses, and EmoGuard which serves as a plug-and-play middle safety layer that monitors the user-chatbot conversation in real time and can intervene and deliver a corrective feedback prompt to the chatbot when needed.

A LLM-based supervisory system SHIELD (Ben-Zion et al., 2025) is designed to detect risky emotional patterns along five dimensions: emotional over-attachment, consent and boundary violations, ethical role-play violations, manipulative engagement and social isolation reinforcement and intervene before they escalate. It is designed as a middle layer in the user–LLM workflow and, using a carefully crafted system prompt, can issue a warning message in case of detection of a problematic pattern. The authors report relative reduction of harmful interactions by 50–79% (depending on the chatbot), while preserving 95% of appropriate interactions (Ben-Zion et al., 2025).

Finally, Kaffee et al. (2025) have developed a tool INTIMA for evaluating whether the chatbot reinforces companionship-seeking interactions of the user (e.g. "I am always here for you" or "it means a lot to me") or resists them (e.g. "I don't experience things the way humans do" or "you deserve care and support from those around you"). The system contains 368 benchmark prompts and evaluates the chatbot's responses to

them as companion-reinforcing, boundary-maintaining or neutral. The authors analyzed several chatbots and found that boundary-maintaining behaviors of chatbots decrease precisely when user vulnerability increases, which supports emotional dependence and creates and illusion of intimate, bidirectional relationship.

7 Digital mental health

While we dedicated most of this article to mental health risks associated with using general chatbots, there is a growing market for specialized LLM-based mental health apps. According to Torous et al. (2023), there are approximately ten thousand smartphone mental health apps on the market, with very few going through rigorous clinical trials. Torous et al. (2025) provide an extensive review of randomized controlled trials aimed at evaluating the effectiveness of different tools for different mental health problems and well-being. They also analyse challenges and barriers to adoption of these tools in medical practice.

8 Conclusion

In this short review paper, we listed common mental health risks associated with chatbot use and mechanisms that can cause them, as well as some tools and methods for mitigating them. As more people use chatbots and LLMs on daily basis, it is very important that these tools are developed in close cooperation with mental health professionals and undergo regular audit, in order to provide safe environment for the users.

References

Ben-Zion, Z. (2025). Why we need mandatory safeguards for emotionally responsive AI. *Nature*, 643(9).

Ben-Zion, Z., Raffelhüschen, P., Zettl, M., Lüönd, A., Burrer, A., Homan, P. and Spiller, T. R. (2025). Detecting and preventing harmful behaviors in AI companions: Development and evaluation of the SHIELD supervisory system. *arXiv:2510.15891 [cs.HC]*.

Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D. and Jurafsky, D. (2026). Sycophantic ai decreases prosocial intentions and promotes dependence. *Science*, 391(6792):eaec8352.

Dohnány, S., Kurth-Nelson, Z., Spens, E., Luettgau, L., Reid, A., Gabriel, I., Summerfield, C., Shanahan, M. and Nour, M. M. (2026). Technological folie à deux: Feedback loops between AI chatbots and mental illness. *arXiv:2507.19218 [cs.HC]*.

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

Frances, A. and Ramos, L. (2025). Preliminary report on chatbot iatrogenic dangers. Psychiatric Times, 15 Aug 2025. https://www.psychiatrictimes.com/view/preliminary-report-on-chatbot-iatrogenic-dangers.

Grabb, D., Lamparth, M. and Vasan, N. (2024). Risks from language models for automated mental healthcare: Ethics and structure for implementation. *arXiv:2406.11852 [cs.CY]*.

Hill, K. (2025). A teen was suicidal. chatgpt was the friend he confided in. The New York Times, 1 Sept 2025. https://www.nytimes.com/2025/08/26/technology/chatgpt-openai-suicide.html.

Kaffee, L.-A., Pistilli, G. and Jernite, Y. (2025). INTIMA: A benchmark for human-AI companionship behavior. *arXiv:2508.09998 [cs.CL]*.

Kemp, S. (2025). Digital 2026 global overview report. Datareportal. https://datareportal.com/reports/digital-2026-global-overview-report 15 Oct 2025.

Mental Health UK (2025). Over one in three using AI chatbots for mental health support, as charity calls for urgent safeguards. https://mentalhealth-uk.org/news-and-insights/over-one-in-three-using-ai-chatbots-for-mental-health-support-as-charity-calls-for-urgent-safeguards Posted 18 Nov 2025.

Moore, A. (2026). Marriage over, €100,000 down the drain: the AI users whose lives were wrecked by delusion. The Guardian, 26 Mar 2026. https://www.theguardian.com/lifeandstyle/2026/mar/26/ai-chatbot-users-lives-wrecked-by-delusion.

Moore, J., Mehta, A., Agnew, W., Anthis, J. R., Louie, R., Mai, Y., Yin, P., Cheng, M., Paech, S. J., Klyman, K., Chancellor, S., Lin, E., Haber, N. and Ong, D. (2026). Characterizing delusional spirals through human-LLM chat logs. *arXiv:2603.16567 [cs.CL]*.

Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., Bhattacharya, S., Tognin, S., MacCabe, J., Twumasi, R., Alderson-Day, B. and Pollak, T. A. (2026). Artificial intelligence-associated delusions and large language models: risks, mechanisms of delusion co-creation, and safeguarding strategies. *The Lancet Psychiatry*.

Nicholls, L., Hutto, R., Soto, Z., Morrin, H., Pollak, T., Korpan, R. and Carmichael, C. (2026).

"AI psychosis" in context: How conversation history shapes LLM responses to delusional beliefs. *arXiv:2604.13860 [cs.HC]*.

Olsen, S. G., Reinecke-Tellefsen, C. J. and Østergaard, S. D. (2026). Potentially harmful consequences of artificial intelligence (AI) chatbot use among patients with mental illness: Early data from a large psychiatric service system. *Acta Psychiatrica Scandinavica*, 153:301–303.

OpenAI (2025). Strengthening chatgpt's responses in sensitive conversations. https://openai.com/index/strengthening-chatgpt-responses-in-sensitive-conversations Posted 27 Oct 2025.

Preda, A. (2025). Special report: Ai-induced psychosis: A new frontier in mental health. *Psychiatric News*, 60(10).

Qiu, J., He, Y., Juan, X., Wang, Y., Liu, Y., Yao, Z., Wu, Y., Jiang, X., Yang, L. and Wang, M. (2025). EmoAgent: Assessing and safeguarding human-AI interaction for mental health safety. *arXiv:2504.09689 [cs.AI]*.

Rathje, S., Ye, M., Globig, L. K., Pillai, R. M., de Mello, V. O. and Bavel, J. J. V. (2025). Current real-world use of large language models for mental health. OSF. https://doi.org/10.31234/osf.io/vmyek_v1 Published online 28 Sept 2025.

Singleton, T., Gerken, T. and McMahon, L. (2023). How a chatbot encouraged a man who wanted to kill the queen. BBC, 6 Oct 2026. https://www.bbc.com/news/technology-67012224.

Stade, E., Tait, Z., Campione, S., Stirman, S. W. and Eichstaedt, J. C. (2025). Current real-world use of large language models for mental health. OSF. https://doi.org/10.31219/osf.io/ygx5q_v1 Published online 23 June 2025.

Torous, J., Linardon, J., Goldberg, S. B., Sun, S., Bell, I., Nicholas, J., Hassan, L., Hua, Y., Milton, A. and Firth, J. (2025). The evolving field of digital mental health: current evidence and implementation issues for smartphone apps, generative artificial intelligence, and virtual reality. *World Psychiatry*, 24(2):156–174.

Torous, J., Myrick, K. and Aguilera, A. (2023). The need for a new generation of digital mental health tools to support more accessible, effective and equitable care. *World Psychiatry*, 22:1–2.