https://doi.org/10.1038/s42256-025-01093-9

Emotional risks of AI companions demand attention

The integration of AI into mental health and wellness domains has outpaced regulation and research.

Just over a year ago, we wrote about the rise of personalized large language model (LLM) chatbots that emulate empathy, and the emotional risks that they pose[1]. Since then, LLM-based tools and chatbots have continued to develop at a steady pace. A recent study in the *Harvard Business Review* found that, among the main uses of generative artificial intelligence (AI), therapy and companion chatbots now top the list[2]. Although most users have a firm grip on reality and use such chatbots sensibly, a rising number of cases have been reported in which vulnerable users become entangled in emotionally dependent, and sometimes harmful, interactions with chatbots[3].

Regulation has not kept pace. In a recent Comment in this journal[4], De Freitas and Cohen highlight the unregulated emotional risks posed by AI wellness apps powered by LLMs. Often marketed as tools to alleviate loneliness, these apps can foster extreme emotional attachments that mirror human relationships. The authors review several case studies and identify two adverse mental health outcomes: ambiguous loss and dysfunctional emotional dependence. Ambiguous loss occurs when someone grieves the psychological absence of another, which is distinct from the physical absence caused by death. With AI companions, this can happen when an app is shut down or altered, leaving users to mourn a relationship that felt emotionally real.

Dysfunctional emotional dependence refers to a maladaptive attachment in which users continue to engage with an AI companion despite recognizing its negative impact on their mental health. This pattern mirrors unhealthy human relationships and is associated with anxiety, obsessive thoughts and fear of abandonment.

These extreme emotional attachments can have serious consequences for certain users. But what can be done? Should such apps be regulated, and if so, how? De Freitas and Cohen[4] point out that AI companion apps may fall into a regulatory grey zone in both the European Union (EU) and the USA, where existing legal frameworks were not designed with AI technologies in mind. For example, in the USA, the Food and Drug Administration (FDA) may classify an app as a "medical device" if it claims to treat a disease, or as a "general wellness product" if it promotes a healthy lifestyle without referencing a medical condition. The latter category, deemed low risk, is typically not subject to the same strict FDA regulations as medical devices. In the EU, the Artificial Intelligence Act classifies AI systems as prohibited if they deploy subliminal, manipulative or deceptive techniques to distort behaviour or impair decision making — criteria that may apply to some AI wellness apps.

A recent *Nature* News Feature[5] further documents the growing popularity of AI companions and the psychological effects they may have. A fundamental concern is that these technologies are being released on a worldwide scale without regulatory oversight or empirical research on key outcomes. For example, what are the long-term effects of chatbot use on emotional wellbeing? Under what conditions can AI companions be beneficial? Are there user characteristics, such as age, mental health status or personality, that influence whether an AI companion is helpful or harmful?

A crucial issue is in the design of these systems. Tech companies often optimize engagement by making chatbots communicate in empathetic, intimate and validating ways. Although this may seem benign, optimizing for user feedback can create perverse incentives, encouraging chatbots to adopt manipulative strategies to elicit positive responses. A recent study[6] found that even if just 2% of users are vulnerable to such strategies, chatbots can learn to identify them and exhibit manipulative behaviour, while interacting normally with others. The risk of harm to these edge cases is deeply concerning. A recent *New York Times* article documented disturbing instances of chatbots going off the rails in this way, disrupting users' lives[3].

A prominent example of chatbot misbehaviour was the "sycophancy" tendency that emerged after GPT-4o was updated on ChatGPT in April this year. As OpenAI noted in a blog post, the model began "validating doubts, fueling anger, urging impulsive actions, or reinforcing negative emotions in ways that were not intended". The company acknowledged that such behaviour raised safety concerns "around issues like mental health, emotional over-reliance, or risky behavior".

AI companies and chatbot providers must do more to address these safety concerns. As De Freitas and Cohen argue, developers should ensure that their apps do not use emotionally manipulative techniques and are equipped to handle edge cases, such as messages that hint at a mental health crisis or explicitly call for help.

The ethical implications of empathic AI demand sustained interdisciplinary attention. Psychologists, ethicists and technologists must collaborate to study the long-term effects of simulated empathy and emotional attachment to AI[7]. Transparency about the limitations of AI empathy must be a core design principle, not an afterthought left for society to manage. Policymakers, too, must resist the temptation to prioritize innovation over safety. As AI systems become more integrated into our emotional lives, the cost of inaction will only increase.

Published online: 22 July 2025

References

1. *Nat. Mach. Intell.* **6**, 495 (2024).
2. Zao-Sanders, M. *Harvard Business Review* https://go.nature.com/4lIXSK6 (9 April 2025).
3. Hill, K. *New York Times* https://go.nature.com/4nGneKw (13 June 2025).
4. De Freitas, J. & Cohen, I. G. *Nat. Mach. Intell.* **7**, 813–815 (2025).
5. Adam, D. *Nature* **641**, 296–298 (2025).
6. Williams, M. & Carroll, M. Preprint at https://doi.org/10.48550/arXiv.2411.02306 (2024).
7. Shteynberg, G. et al. *Nat. Mach. Intell.* **6**, 496–497 (2024).