TYPE Perspective
PUBLISHED 22 May 2025
DOI 10.3389/fmed.2025.1612838

OPEN ACCESS

EDITED BY
Inbar Levkovich,
Tel-Hai College, Israel

REVIEWED BY
Refael Yonatan-Leus,
College of Management Academic Studies,
Israel

*CORRESPONDENCE
Yaakov Ophir
✉ yaakovophir@gmail.com

RECEIVED 16 April 2025
ACCEPTED 05 May 2025
PUBLISHED 22 May 2025
CORRECTED 24 June 2025

CITATION
Ophir Y, Tikochinski R, Elyoseph Z, Efrati Y and Rosenberg H (2025) Balancing promise and concern in AI therapy: a critical perspective on early evidence from the MIT–OpenAI RCT.
Front. Med. 12:1612838.
doi: 10.3389/fmed.2025.1612838

Balancing promise and concern in AI therapy: a critical perspective on early evidence from the MIT–OpenAI RCT

Yaakov Ophir[1,2]*, Refael Tikochinski[1,3], Zohar Elyoseph[4], Yaniv Efrati[5] and Hananel Rosenberg[6]

[1]Department of Education, Ariel University, Ariel, Israel, [2]Centre for Human-Inspired Artificial Intelligence (CHIA), University of Cambridge, Cambridge, United Kingdom, [3]Experimental Psychology Department, University College London, London, United Kingdom, [4]Faculty of Education, University of Haifa, Haifa, Israel, [5]Faculty of Education, Bar-Ilan University, Ramat Gan, Israel, [6]The Moscowitz School of Communication, Ariel University, Ariel, Israel

The emergence of AI therapy chatbots has the potential to reduce the widening gap between the huge demand for psychological support today and the limited availability of professional care. However, some scholars and clinicians are warning that the integration of these chatbots could paradoxically lead to negative outcomes, such as dependence, loneliness, and depression. Recently, a joint research team from MIT and OpenAI conducted a four-week Randomized Controlled Trial (RCT), reporting that "*while participants on average were less lonely after the study… extended daily interactions with AI chatbots can reinforce negative psychosocial outcomes*". Considering the substantial public and academic attention that followed the preprint publication of this RCT, it is crucial to examine the strength of the evidence and the validity of its interpretation before drawing firm conclusions. In this commentary, we offer a careful and appreciative review of this well-designed and timely study. Nonetheless, we argue that due to key analytical limitations, the findings do not substantiate claims of harmful effects. Given the transformative potential of AI-based interventions, we urge caution in interpreting early findings and offer practical guidance for future research.

KEYWORDS

AI therapy, digital mental health, technological panic, mental health innovation, AI in mental health

Introduction

The global burden of mental health disorders remains a pressing concern, particularly in the wake of the COVID-19 crisis (1, 2). Despite decades of research and clinical advancement, rates of mental health diagnoses and psychopharmacological prescriptions continue to rise (3–5). Against this backdrop, many have placed hope in Artificial Intelligence (AI) as a potential game-changer in mental health care.

Psychologically oriented AI chatbots, designed to function as sensitive personal companions, offer a promising avenue to expand access to psychological support—particularly for individuals experiencing mild symptoms or everyday emotional struggles. Such tools might help alleviate some of the pressure on overburdened mental health systems (6).

However, the prospect of "machine-based" psychological assistance also elicits understandable skepticism. Critics argue that AI systems cannot replicate the authentic empathy and human connection that are widely regarded as central to the effectiveness of traditional psychotherapy (7, 8). Others raise concerns about the potential harms of these technologies, including the risk of emotional overdependence or even addictive use, which may reduce real-world social interaction and exacerbate feelings of loneliness (9, 10).

To explore these concerns empirically, a research team from MIT and OpenAI conducted a longitudinal randomized controlled trial (RCT), investigating the psychosocial effects of interacting with AI chatbots over the course of 4 weeks (11). Participants were randomly assigned to engage with chatbots featuring different modalities (text or voice) and conversational styles (open-ended, non-personal, or personal). Psychosocial outcomes—including loneliness, emotional dependence, socialization with real people, and problematic use—were assessed both before and after the study.

The authors present a rich and detailed set of analyses, incorporating multiple interaction effects and subgroup comparisons. As might be expected from such a multifaceted design, the resulting findings are complex and at times challenging to interpret. The discussion section reflects a thoughtful effort to engage with this complexity. Nevertheless, the authors open their discussion with a concise summary statement that appears to crystallize the study's central takeaway—a framing that, in our view, merits closer scrutiny:

> "Our study results show that while participants on average were less lonely after the study—especially after interacting with an engaging voice-based chatbot—extended daily interactions with AI chatbots can reinforce negative psychosocial outcomes such as decreased socialization."

Naturally, these findings drew considerable public and academic interest. Although the study has been shared publicly as a preprint, prior to peer review, its conclusions—particularly regarding potential harms—have already been widely circulated in prominent media outlets such as *Bloomberg* and *Forbes* (12–15). Given the influence such early reports can have on public perception and policy discussions, and considering the broader implications for the development and adoption of AI-based psychological tools, we believe that a careful and balanced examination of the study's methods and interpretations is both timely and warranted.

A close inspection of the MIT–OpenAI RCT

Responding to this need, we now take a closer look at Fang et al.'s RCT. While the study offers a timely and ambitious contribution to the emerging field of AI-based mental health interventions, several aspects of its design and interpretation merit further scrutiny. In the following section, we examine key methodological features of the trial—including the definition of control conditions, the modeling of usage effects, and the interpretation of effect sizes—that complicate strong conclusions about psychosocial harms. Our aim is to clarify what the data do and do not support at this early stage of research.

What counts as a proper placebo?

One of the most fundamental questions in this type of research is: what counts as an appropriate placebo? A traditional waiting-list control group would not suffice here, as it would not allow us to distinguish between the effect of simply interacting with an AI chatbot and the effect of engaging with a chatbot in a personal, emotionally meaningful way. From the structure of the study, it seems the authors recognize this point. However, the chosen reference group in their regressions—the *open-ended* condition—may not serve as an ideal placebo. Because participants in this condition could discuss both personal and non-personal topics, it blurs the line between intervention and control. A stronger internal placebo comparison would be the *non-personal* condition, where conversations were explicitly kept impersonal.

Mixed signals in the reported outcomes

This makes it difficult to interpret the findings with precision. Nonetheless, when we examine Figure 6 and its underlying regression results in Table 6 (Appendix N), a somewhat surprising pattern emerges: there appears to be a slight advantage for the *personal* (i.e., intervention) condition. While the authors report a significantly more moderate decline in socialization in the *non-personal* (placebo-like) group, this advantage is not reflected across the other outcomes. In fact, for loneliness, emotional dependence, and problematic use, the visual trends point to *steeper deterioration* in the *non-personal* condition—even if these effects are not statistically significant in two of the three outcomes. Taken together, Figure 6 does not offer strong support for the idea that the placebo condition is more benign than the intervention; if anything, the opposite appears to be true.

The problem of usage duration and reverse causality

More importantly, the study's attempt to explore the effects of daily usage duration is commendable but methodologically limited. Because *duration was not manipulated* as part of the experimental design, it remains a naturally varying covariate. It is highly plausible—if not likely—that individuals who were more lonely or emotionally vulnerable to begin with were those who used the chatbot more frequently, particularly in the personal condition.

Modeling change: time 2 is not enough

This key methodological issue could have been partially addressed by modeling change scores (i.e., Time 2 minus Time 1) as the dependent variables, rather than modeling "*the final values of loneliness, socialization, emotional dependence, and problematic use measured at week 4…, controlling for their respective initial values measured at the start of the study*" (p. 4). Alternatively, the authors could have retained the longitudinal time variable ("Week")—which appeared in their mixed-effects models in Table 3—in the subsequent regression analyses presented in Tables 4 through 6. These approaches would have enabled a more direct and transparent examination of within-person change over time.

In the absence of such adjustments, a competing interpretation remains: that participants with greater initial distress were simply more inclined to engage with the chatbot more frequently. For example, consider a participant named Jonny, who began the study with the highest possible score on the UCLA Loneliness Scale (which includes eight items scored from 1 to 4). Although Jonny may have experienced a meaningful improvement over the course of the study—say, a reduction from 32 to 22—he would likely still report higher loneliness at Time 2 than other participants. In the current modeling approach, Jonny's high final score would be explained mostly by his baseline loneliness (which had a strong predictive value, $\beta = 0.877^{***}$), yet his daily usage of the chatbot would still be statistically associated with that same Time 2 score, not with his improvement.

In other words, by modeling Time 2 rather than the change from Time 1 to Time 2, the regression cannot distinguish between outcomes caused by chatbot use and pre-existing levels of distress that led to greater use in the first place. This opens the door to reverse causality and renders any interpretation of "dose–response" effects speculative at best.

For this reason, the main takeaway from this study should center on the primary effects reported by the authors—rather than on exploratory interaction terms—namely that "*participants on average were less lonely after the study, especially after interacting with an engaging voice-based chatbot*." This finding, modest as it may be, aligns with long-standing clinical thinking, which views emotional expression, personal disclosure, and the experience of being heard—even in brief or technologically mediated interactions—as potentially beneficial (16, 17). These assumptions have guided psychological theory and practice for decades, and it would be surprising if they were suddenly invalidated simply because the source of empathy was algorithmic.

Moreover, the usage levels in this study were quite low: participants spent on average only 5.32 min per day interacting with the chatbot, and even the most engaged participant averaged less than 28 min daily. Against this backdrop, the idea that such brief, self-directed interactions could meaningfully increase loneliness or foster emotional dependency seems less plausible than the alternative hypothesis—that individuals who were already experiencing distress were more likely to seek out and engage with the chatbot more frequently.

Interpreting the effect sizes

Finally, it is worth considering the magnitude of the reported effects. Across outcomes, the effect sizes in this study were uniformly small—often so small that they are unlikely to translate into meaningful real-world implications. For example, in Table 6 discussed above, the strongest reported interaction effect—the moderating effect of non-personal conversation topics on socialization—was $\beta = 0.05$, which corresponds to an r of approximately 0.05. Other outcomes, such as emotional dependence and problematic use, yielded similar or even smaller coefficients. These findings do not negate the value of the research, but they do urge caution in drawing strong conclusions about practical or clinical significance at this early stage.

Caution in interpreting absence of harm

While our critique has focused on methodological limitations that complicate conclusions about the potential harms of AI-based therapy chatbots, it is equally important to emphasize that these same limitations also restrict any strong inferences about their safety.

First, the four-week duration of the study, although practical for an early-stage trial, is insufficient to capture potential long-term psychosocial outcomes, such as ingrained dependency or sustained reductions in offline socialization. Second, the relatively short average daily usage time of 5.32 min leaves open the possibility that higher or more prolonged exposure could lead to different effects, whether beneficial or adverse. Finally, although our analysis suggests that reverse causality is a plausible explanation for the observed correlation between higher chatbot use and negative outcomes, we do not dismiss the possibility of differential risks. Vulnerable subgroups, such as individuals experiencing significant distress, may respond differently to AI-based interactions and could be disproportionately susceptible to adverse effects.

These caveats highlight the need for caution in interpreting not only the original study by MIT and OpenAI, but also our own critique of its findings. While we challenge strong claims of harm based on the available data, we also acknowledge that the absence of clear evidence for harm should not be mistaken for evidence of safety.

Discussion

The study by Fang et al. raises important and timely questions about the psychological effects of interacting with AI-based therapy chatbots. While the authors should be commended for their rigorous and innovative approach, our analysis suggests that the findings—particularly those related to negative psychosocial outcomes—should be interpreted with caution. Several methodological limitations, including challenges in defining appropriate control conditions and assessing usage effects, complicate strong causal claims and limit the conclusions that can be drawn about both harm and safety.

In our view, the current results do not substantiate concerns about increased loneliness or emotional overdependence. As with every new technology, concerns—both valid and overstated—are to be expected. In previous work, we have examined similar dynamics in the context of screen use, where fears about harm were often amplified despite limited empirical support [e.g., (18, 19)]. To avoid repeating this Sisyphean cycle of technological moral panic (20), we must resist the urge to adopt prematurely alarmist narratives around "AI addiction" (21).

At the same time, we should be careful not to mistake the absence of clear evidence for harm as confirmation of safety. It remains important to stay alert to early signs of overreliance on AI therapy bots, particularly among vulnerable populations. We therefore recommend continued investment in careful, incremental studies—such as the one reviewed here—that can guide responsible, evidence-based implementation.

Future research should include (1) experimental manipulations of usage duration to assess its causal influence on psychosocial outcomes, (2) outcome measures based on changes over time, (3) more clearly defined placebo conditions, and (4) comparison groups receiving conventional psychological treatment. Importantly, even well-established therapeutic modalities may sometimes aggravate symptoms or fail to help certain patients (22). Yet these risks are not a reason to abandon traditional therapy altogether—they are part of the broader reality of mental health care. The same perspective should guide our approach to emerging AI-based interventions.

Rather than expecting perfection from new technologies, we should aim for progress—grounded in evidence, attentive to risks, and open to innovation. In the face of a growing mental health crisis, the question, in our view, should not be whether AI is perfect—but whether we are willing to explore its promise responsibly, with both courage and care.

Data availability statement

The original study reviewed in this commentary is publicly available at: https://www.media.mit.edu/publications/how-ai-and-human-behaviors-shape-psychosocial-effects-of-chatbot-use-a-longitudinal-controlled-study/.

Author contributions

YO: Writing – original draft, Writing – review & editing, Methodology, Investigation, Conceptualization, Validation, Project administration. RT: Writing – review & editing, Investigation, Methodology. ZE: Writing – review & editing, Validation, Investigation. YE: Validation, Investigation, Writing – review & editing. HR: Investigation, Validation, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that Gen AI was used in the creation of this manuscript. Generative AI was used in this manuscript solely for language editing and stylistic refinement. No AI tools were involved in the generation of scientific content, interpretation of findings, or conceptual development.

Correction note

A correction has been made to this article. Details can be found at: 10.3389/fmed.2025.1643202.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Arias D, Saxena S, Verguet S. Quantifying the global burden of mental disorders and their economic value. *EClinicalMedicine*. (2022) 54:101675. doi: 10.1016/j.eclinm.2022.101675

2. Hossain MM, Nesa F, Das J, Aggad R, Tasnim S, Bairwa M, et al. Global burden of mental health problems among children and adolescents during COVID-19 pandemic: an umbrella review. *Psychiatry Res*. (2022) 317:114814. doi: 10.1016/j.psychres.2022.114814

3. APA. (2023). Stress in America™ 2023: a nation grappling with psychological impacts of collective trauma. American Psychological Association (APA). Available online at: https://www.apa.org/news/press/releases/2023/11/psychological-impacts-collective-trauma (Accessed March 5, 2024).

4. Kessing LV, Ziersen SC, Caspi A, Moffitt TE, Andersen PK. Lifetime incidence of treated mental health disorders and psychotropic drug prescriptions and associated socioeconomic functioning. *JAMA Psychiatry*. (2023) 80:1000–8. doi: 10.1001/jamapsychiatry.2023.2206

5. Ormel J, Hollon SD, Kessler RC, Cuijpers P, Monroe SM. More treatment but no less depression: the treatment-prevalence paradox. *Clin Psychol Rev*. (2022) 91:102111. doi: 10.1016/j.cpr.2021.102111

6. Stade EC, Stirman SW, Ungar LH, Boland CL, Schwartz HA, Yaden DB, et al. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *NPJ Mental Health Res*. (2024) 3:12. doi: 10.1038/s44184-024-00056-z

7. Békés V, Aafjes-van Doorn K, Luo X, Prout TA, Hoffman L. Psychotherapists' challenges with online therapy during COVID-19: concerns about connectedness predict therapists' negative view of online therapy and its perceived efficacy over time. *Front Psychol*. (2021) 12:705699. doi: 10.3389/fpsyg.2021.705699

8. Perry A. AI will never convey the essence of human empathy. *Nat Hum Behav*. (2023) 7:1808–9. doi: 10.1038/s41562-023-01675-w

9. Laestadius L, Bishop A, Gonzalez M, Illenčík D, Campos-Castillo C. Too human and not human enough: a grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media Soc*. (2024) 26:5923–41. doi: 10.1177/14614448221142007

10. Salah M., Abdelfattah F., Alhalbusi H., Mukhaini M. A. (2024). Me and my AI bot: exploring the 'AIholic' phenomenon and university Students' dependency on generative AI Chatbots - is this the new academic addiction? Available online at: https://www.researchsquare.com/article/rs-3508563/v2 (Accessed March 26, 2025).

11. Fang C. M., Liu A. R., Danry V., Lee E., Chan S. W. T., Pataranutaporn P., et al. (2025). How AI and human behaviors shape psychosocial effects of Chatbot use: a longitudinal randomized controlled study. Available online at: http://arxiv.org/abs/2503.17473 (Accessed March 26, 2025).

12. Daniel L. (2025). 'ChatGPT is my friend'—OpenAI and MIT study reveals who's most vulnerable to AI attachment. Forbes. Available online at: https://www.forbes.com/sites/larsdaniel/2025/04/01/chatgpt-is-my-friend-openai-and-mit-study-reveals-whos-most-vulnerable-to-ai-attachment/ (Accessed April 16, 2025).

13. Metz R. (2025). OpenAI study finds links between ChatGPT use and loneliness. Bloomberg Available online at: https://www.bloomberg.com/news/articles/2025-03-21/openai-study-finds-links-between-chatgpt-use-and-loneliness

14. Nolan B. (2025). ChatGPT might be making frequent users more lonely, study by OpenAI and MIT media lab suggests. Fortune Available online at: https://fortune.com/2025/03/24/chatgpt-making-frequent-users-more-lonely-study-openai-mit-media-lab/ (Accessed April 16, 2025).

15. Varanasi L. (2025). The hidden cost of brainstorming with ChatGPT. Bus Insid Available online at: https://www.businessinsider.com/openai-chatgpt-brainstorming-addiction-dependence-negative-consequences-mit-research-2025-3 (Accessed April 16, 2025)

16. Kennedy-Moore E, Watson JC. How and when does emotional expression help? *Rev Gen Psychol*. (2001) 5:187–212. doi: 10.1037/1089-2680.5.3.187

17. Roos CA, Postmes T, Koudenburg N. Feeling heard: operationalizing a key concept for social relations. *PLoS One*. (2023) 18:e0292865. doi: 10.1371/journal.pone.0292865

18. Ophir Y, Rosenberg H, Tikochinski R. What are the psychological impacts of children's screen use? A critical review and meta-analysis of the literature underlying the World Health Organization guidelines. *Comput Hum Behav*. (2021) 124:106925. doi: 10.1016/j.chb.2021.106925

19. Ophir Y, Rosenberg H, Tikochinski R, Dalyot S, Lipshits-Braziler Y. Screen time and autism spectrum disorder: a systematic review and meta-analysis. *JAMA Netw Open*. (2023) 6:e2346775. doi: 10.1001/jamanetworkopen.2023.46775

20. Orben A. The Sisyphean cycle of technology panics. *Perspect Psychol Sci*. (2020) 15:1143–57. doi: 10.1177/1745691620919372

21. Ciudad-Fernández V, von Hammerstein C, Billieux J. People are not becoming "AIholic": questioning the "ChatGPT addiction" construct. *Addict Behav*. (2025) 166:108325. doi: 10.1016/j.addbeh.2025.108325

22. Lambert MJ. What have we learned about treatment failure in empirically supported treatments? Some suggestions for practice. *Cogn Behav Pract*. (2011) 18:413–20. doi: 10.1016/j.cbpra.2011.02.002