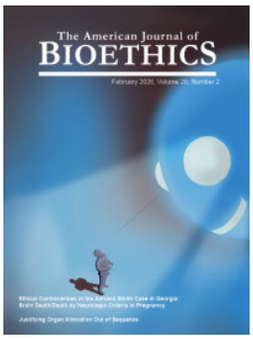

The American Journal of Bioethics

ISSN: 1526-5161 (Print) 1536-0075 (Online) Journal homepage: www.tandfonline.com/journals/uajb20

The Potential Harms of AI Psychotherapy: A Fear as Old as ELIZA

Robert Ranisch & Lukas J. Meier

To cite this article: Robert Ranisch & Lukas J. Meier (2026) The Potential Harms of AI Psychotherapy: A Fear as Old as ELIZA, The American Journal of Bioethics, 26:2, 69-71, DOI: 10.1080/15265161.2025.2608640

To link to this article: https://doi.org/10.1080/15265161.2025.2608640

THE AMERICAN JOURNAL OF BIOETHICS
2026, VOL. 26, NO. 2, 69–71
https://doi.org/10.1080/15265161.2025.2608640

Taylor & Francis
Taylor & Francis Group

OPEN PEER COMMENTARIES

OPEN ACCESS Check for updates

The Potential Harms of AI Psychotherapy: A Fear as Old as ELIZA

Robert Ranisch[a] and Lukas J. Meier[a,b]

[a]University of Potsdam; [b]Harvard University

"We have written a computer program which can conduct psychotherapeutic dialogue" (Colby et al. 1966, 148). What sounds like a statement about the latest generative-AI bot was, in fact, formulated sixty years before the current boom in machine intelligence. In 1966, Stanford psychiatrist Kenneth M. Colby and his colleagues published one of the earliest reports describing a computer-based system for psychotherapy. It was the same year in which, at MIT, computer-science pioneer Joseph Weizenbaum introduced his famous natural-language processing program ELIZA (Weizenbaum 1966), which many regard as the world's first chatbot.

The two concurrent publications marked the starting points of a fierce controversy about the moral limits of automating psychotherapy—a debate that went on for several decades but whose ethical significance, ironically, has now been largely forgotten within contemporary bioethics. Revisiting this early exchange is interesting not just from a historical perspective. As Tekin and Delehanty's Target Article shows (Tekin and Delehanty 2026), many motifs of that initial controversy resurface today, yet their origins are not widely known. Revisiting the early debate can therefore inform discussions on the boundaries of AI-based therapy at a time when many of the same questions are suddenly reasserting themselves. This is what we intend to do in this short comment.

When, in 1964, Weizenbaum began developing what would become the ELIZA program, his goal was to explore the computational mechanisms required to simulate dialogue. Although Weizenbaum adjusted the system to mimic Rogerian psychotherapy (the now-famous DOCTOR script), he had no ambition to actually pioneer computer-based therapy. On the contrary, he later described the script as a "parody" of nondirective psychotherapy, chosen precisely because such conversations require no contextual or real-world understanding. (Weizenbaum 1976, 188–9). Thus, the point that Weizenbaum set out to demonstrate seems to be the superficiality of human-machine communication (Zeavin 2021). Computerizing clinical practice was not among his goals. The latter, however, was exactly what Colby, who was acquainted with Weizenbaum, intended to accomplish: in the very same year that saw the unveiling of ELIZA, Colby, together with colleagues, published a pioneering article on the first attempts at computerizing psychotherapy (Colby et al. 1966). Later, he became known for developing a natural-language therapeutic learning program delivering cognitive behavioral therapy for depression.

Weizenbaum repeatedly voiced his "shock" that psychiatrists—and especially his former friend Colby—were seriously pursuing computer-based psychotherapy, although Colby later offered a different account of these events (Colby 1999a). In his *Computer Power and Human Reason*, Weizenbaum even leveraged Colby's aspirations as a central cautionary tale. While he had little doubt that psychotherapy chatbots were technically feasible, he categorically rejected their use, regarding them as "immoral" (Weizenbaum 1976, 269). Colby, in turn, dismissed Weizenbaum's stance as "pseudo-moralizing" and argued that *foregoing* the opportunities that the novel technology will bring would be an immoral course of action (Colby in: Schank 1976, 10).

This initial disagreement evolved into a rich academic and at times personal dispute, eventually extending far beyond the two individuals (Schank 1976; Zeavin 2021). Despite its polemical undertones, the debate anticipates many core themes of today's discussions about the potential harms and benefits of AI-mediated psychotherapy. Tekin and Delehanty—apparently without being aware of the historical background—reproduce several of its core argumentative patterns. They begin by stressing the persistent lack of access to psychotherapy, aligning with the pragmatist view that harm should be assessed in terms of

CONTACT Robert Ranisch ranisch@uni-potsdam.de Faculty of Health Sciences Brandenburg, University of Potsdam, Am Mühlenberg 9, Potsdam, Brandenburg, 14476 Germany.

outcomes and that AI-based tools could help fill a treatment gap. Likewise, Colby, emphasizing that the need for mental healthcare "will always outstrip the manpower" that human therapists could provide (Colby in: Schank 1976, 10), envisioned a future in which "several hundred patients an hour" (Colby et al. 1966, 152) might be treated by computer systems.

Later in their paper, however, Tekin and Delehanty express concerns that the ubiquity of automated psychotherapeutic tools may induce problematic dependencies in their users, echoing one of Weizenbaum's replies to Colby. While Weizenbaum never tested ELIZA's impact on human patients in an empirically sound manner, he reported anecdotal evidence of users forming emotional attachments to the program, which strongly shaped his objections (Weizenbaum 1976, 189). Fascinatingly, long before the rise of generative AI, his reservations appear rather prophetic: Weizenbaum already recognized an "aura" surrounding advanced technologies and observed people's tendency to project human traits onto simple computational systems—a phenomenon later conceptualized as the "ELIZA effect" (Zeavin 2021). He even warned that brief interactions with such systems might "induce powerful delusional thinking in quite normal people" (Weizenbaum 1976, 7), seemingly anticipating present-day observations of similar effects and the resulting discussions about "chatbot psychosis" (Fieldhouse 2025).

At the core of Weizenbaum's critique, however, lay the conviction that "there are some human functions for which computers *ought* not to be substituted" (Weizenbaum 1976, 270). Thus, for Weizenbaum, the fundamental issue was not what ELIZA, or similar systems, could or could not do from a technological point of view; nor whether they might in fact help patients. The problem, he believed, was that the systems *are machines*.

Relatedly, Tekin and Delehanty wonder whether the capacity for self-reflection, which they take to be an essentially human trait necessarily absent in psychotherapeutic chatbots, may result in poorer performance. However, in gesturing toward an *instrumental* disadvantage as the main consequence of the lack of this feature, they now take sides with Weizenbaum's critics. Some commentators suggest that Weizenbaum conflated the question of whether technology can replicate human capabilities with the question of whether doing so would be desirable (Lem 1980). Cars or airplanes, after all, do not work by mimicking legs or wings. Their purpose is not to replicate human or animal movement but to achieve certain *outcomes*. By analogy, chatbots need not understand, reflect, or empathize like humans. As Colby argues, they must only "deliver the requisite simulated conversational goods" (Colby 1999b, 7). The central question, Colby maintains, is "whether or not communication with such a device can benefit a person suffering from mental disorder" (Colby et al. 1966, 152).

If chatbots were to become successful at establishing therapeutic alliances—the strong interpersonal bonds between therapists and patients that facilitate the aims of psychotherapy—Tekin and Delehanty would, it seems, not object to their deployment. This sets them apart from Weizenbaum, who saw principled reasons for requiring humans in therapeutic relationships: one should not, he submitted, "substitute a computer system for a human function that involves interpersonal respect, understanding and love" (Weizenbaum 1976, 269).

In summary, two enduring positions were already crystallizing by the middle of the past century. On one side, there is what one may term Weizenbaum's *human exceptionalism*: a focus on which human characteristics machines lack, combined with an emphasis on the therapeutic *process*, i.e., the relational and personal qualities required for care. For Weizenbaum, even if a machine could facilitate the desired medical outcome, it *ought not* to replace human therapists, because doing so would violate essential values.

On the other side of the spectrum sits Colby's *computational pragmatism*: a focus on machines' many promises, an emphasis on therapeutic *outcomes*, and a commitment to leveraging technology to support underserved populations. From this perspective, mechanisms matter less than effects. If technology can help, it should be developed and used.

As this brief comparison with the argumentative structure of the Target Article indicates, the original positions from the pioneers of natural-language processing have not lost any of their significance. Sixty years into the future, machines are, once again, being considered as replacements for psychotherapists. While the technology has changed, the arguments remain the same.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

FUNDING

The work was supported by the VolkswagenStiftung as part of the Digital Medical Ethics Network (grant number 9B233).

ORCID

Robert Ranisch http://orcid.org/0000-0002-1676-1694
Lukas J. Meier http://orcid.org/0000-0002-3316-3928

REFERENCES

Colby KM. 1999a. Dialogue programs I have known and loved over 33 years. In: Wilks Y, editor. Machine conversations. Springer US, p. 1–3. https://doi.org/10.1007/978-1-4757-5687-6.

Colby KM. 1999b. Comments on human–computer conversation. In: Wilks Y, editor. Machine conversations. Springer US, p. 5–8. https://doi.org/10.1007/978-1-4757-5687-6

Colby KM, Watt JB, Gilbert JP. 1966. A computer method of psychotherapy: preliminary communication. J Nerv Ment Dis. 142(2):148–152. https://doi.org/10.1097/00005053-196602000-00005

Fieldhouse R. 2025. Can AI chatbots trigger psychosis? What the science says. Nature. 646(8083):18–19. https://doi.org/10.1038/d41586-025-03020-9

Lem S. 1980. On science, pseudo-science, and some science fiction (F. Rottensteiner, Trans.). Science Fiction Studies. 7(Part 3):330–338. https://doi.org/10.1525/sfs.7.3.0330

Schank RC. 1976. The Weizenbaum controversy. SIGART Bull. 59:7–11. https://doi.org/10.1145/1045270.1045271

Tekin Ş, Delehanty M. 2026. Beyond doomsday fears: why we need to consider the potential harms of AI psychotherapy. Am J Bioeth. 26(2):45–55 https://doi.org/10.1080/15265161.2025.2457724

Weizenbaum J. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. Commun ACM. 9(1):36–45. https://doi.org/10.1145/365153.365168

Weizenbaum J. 1976. Computer power and human reason: from judgement to calculation. W. H. Freeman & Company.

Zeavin H. 2021. The distance cure: a history of teletherapy. The MIT Press.

THE AMERICAN JOURNAL OF BIOETHICS
2026, VOL. 26, NO. 2, 71–75
https://doi.org/10.1080/15265161.2025.2608628

OPEN PEER COMMENTARIES

Against What Standard? Why AI Therapists Face Impossible Expectations

Haneen Abu Ghanem and Dov Greenbaum

Reichman University

Tekin and Delehanty (2026) provide a valuable systematic framework for assessing harms in digital mental health. However, to apply this framework effectively, regulators must also evaluate AI mental health tools by comparative net harm relative to actually available alternatives—including waitlists or the absence of treatment—rather than against idealized human therapy that many never receive. Systemic biases currently distort that evaluation. We identify these biases, trace their institutionalization, and propose alternatives. Our aim is to extend their harm framework by embedding it within a comparative net-harm analysis that also accounts for the risks of nondeployment and persistent access gaps. We argue that current regulatory frameworks systematically mismeasure AI mental health risk by ignoring the harms of nondeployment and unmet need.

TWO CORE EVALUATION BIASES

Idealized comparator bias evaluates AI against skilled clinicians and optimal care contexts unavailable to most patients. Tekin and Delehanty frame the choice as AI versus competent human therapists—yet for underserved populations, this comparison obscures the actual question: what is the realistic baseline when standard care is inaccessible? As Tekin (2021) argues in the context of digital phenotyping, idealized baselines obscure real-world constraints and can amplify misclassification harms by creating expectations of clinical perfection that neither humans nor AI meet. This perfection asymmetry bias tolerates human imperfections while demanding AI perfection. Human therapists bring cognitive biases, inconsistent reasoning, and knowledge gaps (Tversky and Kahneman 1974)—yet we license them without transparency or systematic outcome tracking.

CONTACT Dov Greenbaum dov.greenbaum@runi.ac.il Reichman University, Herzliya, Israel.