


OPEN PEER COMMENTARIES



Conversational Artificial Intelligence and Distortions of the Psychotherapeutic Frame: Issues of Boundaries, Responsibility, and Industry Interests

Meghana Kasturi Vagwala^a  and Rachel Asher^b^aHarvard Medical School; ^bBrigham and Women's Hospital

Sedlakova and Trachsel argue that conversational artificial intelligence (CAI) is more than a mere tool, but not quite an agent, as it “simulates having a therapeutic conversation [but] does not really have it” (Sedlakova and Trachsel 2023, 9). The authors emphasize that a failure to understand this distinction “endangers the autonomy and psychological integrity of users” (Sedlakova and Trachsel 2023, 10). Yet, developers of CAI explicitly aim to replicate human experiences of therapeutic alliance among their user base (Beatty et al. 2022; Darcy et al. 2021). In doing so, companies funding the creation of CAI technologies produce conditions in which the line between real and simulated therapy is perhaps intentionally blurred. So too are the expected boundaries between therapist and client, with websites for two of the most popular CAI applications, Woebot and Wysa, advertising access that is “on-demand” and “whenever you want” (“Woebot Health” n.d.; “Wysa—Everyday Mental Health” n.d.) Expanding upon Sedlakova and Trachsel’s critique of CAI’s simulation of therapeutic conversation and therefore simulation of therapeutic alliance, we argue that CAI distorts therapeutic frame, which further problematizes CAI developers’ claims of alliance and raises ethical concerns regarding the limits of clinical benefit CAI can provide users.

Therapeutic frame is a concept in psychotherapy that is composed of relational factors, such as the therapist and patient’s expectations of one another, as well as environmental ones, such as fixed times and durations for sessions, privacy, and confidentiality. These environmental and relational factors give meaning and foundational structure to the therapeutic alliance. Unlike human therapists who are rational agents with the ability to understand and apply concepts,

CAI does not develop the broader psychiatric formulations necessary for understanding which approaches to therapeutic frame best serve which patients. The clinician’s fundamental task is to figure out how to treat a patient. To this end, the formulation acts as the compass, as it synthesizes together a unique profile for each patient that considers predisposing, precipitating, perpetuating, and protective factors alongside biological, psychological, and social dynamics that could explain symptoms, clarify diagnosis, and guide the path of treatment. CAI at present, whether because its developers cannot yet program formulation or do not see a need for formulation in the business model for CAI, takes a more one size fits all approach, providing instantaneous response and 24/7 availability as desirable features for all users.

The therapeutic frame promises continuity but not on-demand access, a dynamic that psychologist Anne Gray represents “the reality principle - the fact that we cannot have all that we desire instantaneously” (Gray 2013). In contrast to enticing potential users with the message that “there are no such things as waiting rooms or appointments here” (“Woebot Health” n.d.), the therapeutic frame is predicated on the notion that the frustration of waiting between appointments for some patients is an important part of the therapeutic work, allowing them space to individuate and learn to contain their own emotions. For other patients, particularly those who have more serious safety concerns for self-harm, selective employment of more frequent touchpoints with a clinician can be appropriate and productive aspects of the therapeutic frame. Human therapists can formulate, contextualize the frequency of communication, consider its potential relationship to clinical severity, and

leverage this knowledge to improve patients' insight into these behaviors. CAI, however, cannot "provide robust and complex explanations that might help individual users to better understand their very individual experiences" (Sedlakova and Trachsel 2023, 9). To our knowledge, CAI does not analyze the frequency of contact to explicitly assess severity of clinical condition, rather focusing its detection of severity on more obvious markers such as use of certain phrases with concerning wording (Tekin 2021).

Despite this fundamental reorientation of frame away from what is indicated by psychotherapy and toward businesses' interest in establishing customer satisfaction, developers of CAI assert that their products are based on "proven therapies." It is difficult to analyze how the AI algorithms carry out the work of therapy, given that these algorithms are proprietary and inaccessible to scrutiny—the so-called "black-box problem" referenced by the target article (Sedlakova and Trachsel 2023). However, websites for popular CAI products depict advertising that blends general features of therapy, such as emotional intelligence and capacity to establish therapeutic alliance, together with specific references to "evidence-based" modalities, particularly Cognitive Behavioral Therapy and Dialectical Behavioral Therapy ("Wysa—Everyday Mental Health" n.d.; "Woebot Health" n.d.). Yet, by cherry-picking "appealing" features of therapy while leaving out the relational and environmental factors that contain individuals' development of self and understanding of relationships that occurs in therapy, CAI developers distort the therapeutic frame so foundational to the evidence-based techniques they claim to employ. Such cherry-picking is not without cost for vulnerable populations: app developers' portrayal of their algorithms, as emotionally intelligent, validating conversational partners may prove counterproductive for users with psychotic disorders, for whom validation might reinforce delusions. For individuals with borderline personality disorder, a psychiatric condition characterized by unstable self-image, instability in relationships, and impulsivity, DBT is a gold standard treatment and relies on the therapeutic frame to teach patients to tolerate boundaries and apply the aforementioned skills. The contradiction raised by CAI that advertises both 24/7 access and DBT may obstruct opportunities to effectively engage in the realistic constraints of interpersonal relationships, tolerate distress, and practice self-efficacy through independently navigating life challenges.

Developers of CAI may rebut these arguments by stating that their claims of efficacy are based on

research only with individuals identified as having depression or anxiety. While individuals with severe psychiatric conditions are often explicitly excluded in CAI research, they likely are part of a largely unrestricted user base (Kretzschmar et al. 2019). Patients with severe personality disorders or psychotic disorders are often undiagnosed and often do score above threshold on the Patient Health Questionnaire and Generalized Anxiety Disorder Assessment (Fowler et al. 2018), screening measures of depression and anxiety frequently used by Woebot and Wysa in their research trials (Inkster, Sarda, and Subramanian 2018; Fitzpatrick, Darcy, and Vierhile 2017). Equally as significant is the likelihood that the nuances of harms such as diminished self-efficacy, distress tolerance, or delusion reinforcement (which could be assessed in formulation by a human therapist), would be difficult to capture unless specific interview questions or psychometrics assessing these constructs were used as outcomes measures in research.

While distortions of the therapeutic frame raise numerous ethical concerns, some users may indeed have the ability to appropriately self-regulate and use features of CAI in the service of building self-efficacy, practicing mindfulness, and furthering their overall wellness. We can envision CAI's flexibility and approachability as an accessible entry point to therapeutic principles and skills, including in less well-resourced areas where therapy may be less accessible or cost-prohibitive (although potential messaging regarding therapy for less well-resourced individuals being pharmed out to bots rather than humans is, to put it mildly, problematic). Perhaps the technological sophistication embedded in CAI that allows users to feel a sense of an alliance can best be harnessed through the avenues of listening rather than speaking, detecting rather than therapizing, interacting with users to recognize digital phenotypes of mental illness that human clinicians can then follow up on in a timely fashion. As we continue to gain a better understanding of this "novel type of epistemic exchange" between CAI and user (Sedlakova and Trachsel 2022, 8), it will be crucial for us to also consider what novel framework and new form(s) of alliance CAI may be introducing to the field of psychology. The benefits of CAI for users likely lie in the yet uncharted landscape of human-AI relationships, and not in simulations of psychotherapy. The benefits of CAI for companies are far less ephemeral: A large user base whose time and attention are increasingly occupied by and satisfied with the ease and conveniences of human-bot interaction. Perhaps in the future, the implications of this

will be far less unsettling to most, if unsettling at all, and the bristling sensation of artifice in a clinician's response to companies' claims of bot-user therapeutic alliance would be less warranted.

Beyond (although related to) matters of ethics, we must ask of ourselves, is this something that we want? Do we want to orchestrate algorithmic encapsulation of fundamental components of alliance, such as empathy, as advertised on app websites ("Woebot Health" [n.d.](#); "Wysa—Everyday Mental Health" [n.d.](#))? Perhaps computational modeling could facilitate instruction of these valuable skills to clinicians. In psychotherapy and in human life, empathy can be conceived as notoriously difficult to teach, skilled yet spontaneous, deeply inspiring, error-prone and limited, and particularly beautiful in its mysterious framing of human connection and conscious experience. Historically, however, phenomena with almost magical qualities are often relegated to the realm of reasoned scientific explanation over time. If we do prove or convince ourselves that we can fully model empathy or consciousness, will we miss the unpredictability and awe of the unknown? Perhaps by then the cultural memory of a time when we could, and this even mattered, will have long-since faded.

FUNDING

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Meghana Kasturi Vagwala  <http://orcid.org/0000-0001-8244-768X>

REFERENCES

Beatty, C., T. Malik, S. Meheli, and C. Sinha. 2022. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): A mixed-methods study.

- Frontiers in Digital Health* 4:847991. doi:[10.3389/fdgth.2022.847991](https://doi.org/10.3389/fdgth.2022.847991).
- Darcy, A., J. Daniels, D. Salinger, P. Wicks, and A. Robinson. 2021. Evidence of human-level bonds established with a digital conversational agent: Cross-sectional, retrospective observational study. *JMIR Formative Research* 5 (5):e27868. doi:[10.2196/27868](https://doi.org/10.2196/27868).
- Fitzpatrick, K. K., A. Darcy, and M. Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health* 4 (2):e19. doi:[10.2196/mental.7785](https://doi.org/10.2196/mental.7785).
- Fowler, J. C., J. D. Clapp, A. Madan, J. G. Allen, B. Christopher Frueh, P. Fonagy, and J. M. Oldham. 2018. A naturalistic longitudinal study of extended inpatient treatment for adults with borderline personality disorder: An examination of treatment response, remission and deterioration. *Journal of Affective Disorders* 235 (August): 323–31. doi:[10.1016/j.jad.2017.12.054](https://doi.org/10.1016/j.jad.2017.12.054).
- Gray, A. 2013. *An introduction to the therapeutic frame*. London, UK: Routledge.
- Inkster, B., S. Sarda, and V. Subramanian. 2018. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR-MHealth and UHealth* 6 (11):e12106. doi:[10.2196/12106](https://doi.org/10.2196/12106).
- Kretzschmar, K., H. Tyroll, G. Pavarini, A. Manzini, and I. Singh. 2019. Can your phone be your therapist? Young people's ethical perspectives on the use of fully automated conversational agents (Chatbots) in mental health support. *Biomedical Informatics Insights* 11 (March): 1178222619829083. doi:[10.1177/1178222619829083](https://doi.org/10.1177/1178222619829083).
- Sedlakova, J., and M. Trachsel. 2023. Conversational artificial intelligence in psychotherapy: A new therapeutic tool or agent? *The American Journal of Bioethics* 23 (5):4–13. doi:[10.1080/15265161.2022.2048739](https://doi.org/10.1080/15265161.2022.2048739).
- Tekin, Ş. 2021. Is big data the new stethoscope? Perils of digital phenotyping to address mental illness. *Philosophy & Technology* 34 (3):447–61. doi:[10.1007/s13347-020-00395-7](https://doi.org/10.1007/s13347-020-00395-7).
- "Woebot Health." [n.d.](#) Woebot Health. Accessed January 27, 2023. <https://woebothealth.com/>.
- "Wysa – Everyday Mental Health." [n.d.](#) Wysa – everyday mental health. Accessed January 27, 2023. <https://www.wysa.com/>.