Serving Whom? Ethical and Practical Limits of AI Mental Health Chatbots for Marginalized Communities

Omotunde Falade, Stanford University

Abstract

AI mental health chatbots could offer a promising solution to the care gap faced by underserved communities, especially during times of social isolation and health crisis. These tools provide round-the-clock, low-cost, and stigma-free support. Yet, this paper explores how current implementations face challenges that may limit their long-term efficacy and equitable impact. Drawing on recent empirical studies and ethical scholarship, I evaluate the capabilities and constraints of AI chatbots as short-term mental health supports and propose safeguards that promote inclusive, culturally relevant, and ethically responsible deployment. This study presents new empirical insights from a qualitative interview study with 18 low-income, first-generation community college students of color, whose reflections on chatbot use underscore the importance of trust, cultural resonance, and long-term engagement. Rather than comparing or critiquing specific products, this research centers community voices and advocates for collaborative, community-informed improvement. It also draws from an ethically sourced and protected dataset created by Dr. Harriett Jernigan at Stanford University, which provides a compelling counterexample of culturally specific chatbot design.

Introduction

In recent years, mental health concerns have risen sharply, particularly among young people, marginalized communities, and those living under economic precarity (Pozzi & De Proost, 2024). Professional therapy remains costly and scarce, with long waitlists and a national shortage of clinicians. As a result, AI-powered mental health chatbots have emerged as a potential solution. With the promise of affordability, 24/7 availability, and non-judgmental support, tools like Woebot, Wysa, and Replika offer therapeutic conversations to millions (Casu et al., 2024). Yet, the same qualities that make these tools appealing—automation, scalability, and anonymity—can render them less effective when users are in crisis or require cultural attunement. This paper investigates both the promise and limitations of these tools and proposes that their development be guided by collaborative partnerships with the communities they aim to serve.

Literature Review

For marginalized populations, as well as people living in remote areas or those uncomfortable discussing mental health in person, chatbots may represent

their only consistent point of contact. New tools are continuously appearing on the market, such as *InnerPeaceAI*, introduced at the 2025 IEEE Conference on Multi-Agent Systems for Collaborative Intelligence, exemplify this trend by offering real-time, AI-driven therapy sessions tailored to individual needs. Built on natural language processing and machine learning, InnerPeaceAI promises on-demand support without the barriers of time, location, or cost, particularly targeting those who cannot afford or access conventional therapy. These platforms aim to bridge gaps in care for individuals in low-income or rural communities and claim to enhance well-being by offering confidential, affordable, and user-friendly mental health care (Revathi et al., 2025). Moreover, chatbots offer some emotional support by simulating active listening techniques such as paraphrasing and affirmations. They are also increasingly trained in evidence-based methods such as cognitive behavioral therapy (CBT), which can guide users through mood tracking, journaling, and thought reframing (Eryılmaz & Başal, 2024). Yet despite these advantages, growing evidence shows that chatbots often fall short of their potential in practice. One study suggests that bots may offer responses comparable to human therapists, Ho, Hancock, and Miner (2018). Using a Wizard-of-Oz design, they found that participants who believed they were conversing with a chatbot experienced emotional, psychological, and relational benefits that were statistically indistinguishable from those who believed they were speaking to a human. Emotional disclosures, regardless of partner identity, led to greater perceived understanding, deeper intimacy, and more cognitive reappraisal than factual disclosures.

While the appeal of AI chatbots lies in their accessibility and ability to simulate therapeutic interactions, other empirical studies have begun to question their actual impact. In a mixed-method systematic review, Gaffney, Mansell, and Tai (2019) analyzed 13 studies evaluating 11 different conversational agents and found that all studies reported reductions in psychological distress following chatbot interventions. Five controlled studies even demonstrated significant improvements over inactive controls, suggesting these tools can help alleviate mild to moderate symptoms of depression, anxiety, and loneliness. However, the review also emphasized significant methodological limitations, including small sample sizes, self-selection bias, and lack of long-term follow-up. Users appreciated the bots flexibility, accessibility, and ability to simulate empathy, but also expressed frustration with limited understanding, repetitive content, and shallow conversation depth. These findings highlight the importance of improving design, evaluating long-term outcomes, and considering safety risks, particularly in handling high-risk situations like suicidal ideation.

With similar findings, Limpanopparat, Gibson, and Harris (2024) synthesized data from 49 studies involving 20 mental health chatbots and 14,594 participants to better understand how user perceptions of these tools shape their

adoption and effectiveness. While users generally viewed chatbots as accessible and helpful for managing symptoms of anxiety and depression, persistent concerns about privacy, personalization, and the chatbot's ability to authentically simulate human conversation limited long-term engagement. The authors argue that chatbot design must prioritize user trust and emotional resonance, especially by tailoring content and delivery to the cultural and psychological needs of target populations. Despite moderate evidence of clinical benefit, the review underscores the need for more longitudinal studies with diverse samples to ensure equitable outcomes in real-world settings.

A 2025 review by Yuan et al. offers the most expansive synthesis to date of large language model (LLM)-based mental health chatbots, identifying over 50 such tools, including 22 specifically targeting conditions like depression, anxiety, and suicide ideation. Their findings reinforce existing concerns about privacy, over-dependency, and inappropriate crisis responses, even with innovations such as Retrieval-Augmented Generation (RAG), instruction tuning with CBT principles, and participatory design involving real patients and clinicians. Importantly, the review notes that few chatbots have been evaluated rigorously or tailored for vulnerable populations—especially in workplace or organizational contexts, where mental health needs are rising. This study supports my central claim: chatbots may offer short-term support, but without careful oversight, cultural relevance, and integration with human care, their benefits will remain limited.

They also appear to be a short-term intervention without durable support. A systematic review by Fleming et al. (2018) reveals a stark contrast in real-world implementation. Analyzing 11 studies of self-help digital tools for mood and anxiety disorders, they found real-world engagement rates were drastically lower than in clinical settings. Although some interventions reached tens of thousands of users, completion rates ranged from as low as 0.5% to 28.6%, and very few users sustained engagement over time. The review highlighted inconsistencies in how engagement is measured and reported, making comparisons difficult but underscoring a key issue: interventions that appear effective in controlled environments often fail to maintain user engagement once deployed to the public. These findings raise concerns about the scalability of AI-based mental health tools. If people from underserved or skeptical communities are already wary of these tools, the drop-off in engagement in broader populations, even among self-selected users, suggests we need deeper attention to user trust, cultural fit, and sustained motivation.

Many AI mental health tools are designed without meaningful input from the communities most affected by mental health inequities. As Pozzi and De Proost (2024) emphasize, when marginalized populations are excluded from design processes, technologies may reproduce the very structural inequities they claim to alleviate. Tawiah and Monestime (2024) underscore

the urgent need for participatory design models—ones that incorporate the lived expertise of racial and ethnic minorities, low-income communities, and non-English speakers into every stage of development.

> If the development team lacks diversity or fails to consider the specific needs of various populations, the resulting tools may inadvertently reflect the biases of their creators. This can lead to tools that are less effective or even harmful for certain groups, exacerbating existing disparities in mental health care. (Tawiah & Monestime, 2024)

These studies lend empirical support to the "computers as social actors" (CASA) framework, which posits that people respond to chatbots in socially and emotionally similar ways as they do to humans. Achieving this kind of engagement in real-world settings requires deliberate design to simulate warmth, understanding, and validation—elements that are often missing from commercially deployed chatbots. The limitations students experienced, then, are not necessarily the fault of any individual product, but instead reflect broader design gaps in AI mental health tools and the limitations of the data on which they are built.

A major contributor to these limitations is the quality and scope of training data. The seminal paper "On the Dangers of Stochastic Parrots: Can language models be too big? 🦜" (Bender, Gebru, McMillan-Major, & Mitchell, 2021) brought critical attention to the ethical and technical risks of training large language models on massive, indiscriminately scraped datasets from the internet. These models, the authors argue, do not "understand" language, but remix and regurgitate it in statistically plausible ways. As a result, they often encode harmful biases, reproduce stereotypes, and amplify misinformation—especially when marginalized voices are underrepresented or misrepresented in the data. The paper also raised concerns about the environmental cost of training large models, the opacity of their development, and the tendency to scale systems without clear purpose or accountability.

Since that landmark paper, additional studies have reinforced these concerns. Scholars such as Birhane (2021), Bommasani et al. (2021), Paullada et al. (2021), and Liang et al. (2022) have demonstrated that flawed training data can reinforce structural inequities in AI applications, including in healthcare. These concerns are heightened in mental health settings, where AI models are particularly prone to so-called "hallucinations"—instances where a model generates false or fabricated information that appears coherent or plausible. While the term is borrowed from human psychology, it misleadingly suggests intentionality or subjective experience, when in fact these errors are the result of statistical pattern-matching without understanding. In mental health contexts, such hallucinations are particularly dangerous because users may be emotionally vulnerable and inclined to trust the chatbot's responses as credible

guidance, increasing the risk of misinformation, misdiagnosis, or even harm during moments of crisis (Huang, 2025 Aljamaan, et al, 2024). With companies continuing to market these tools as universally applicable, despite their limitations, researchers like Raji et al. (2021) have shown how such tools often fail to serve racial and ethnic minorities, low-income populations, and non-English speakers whose needs were not accounted for in the development phase.

Recent work in medicine further affirms these risks. Thirunavukarasu et al. (2023), in a comprehensive *Nature Medicine* review, argue that while large language models like ChatGPT show promise in clinical support settings, they remain unreliable for autonomous medical use due to their tendency to produce false but persuasive content. These flaws are especially dangerous in mental health contexts, where users may be vulnerable and less equipped to evaluate misinformation. Without stronger ethical guidelines, regulatory oversight, and inclusive design practices, even the most advanced AI tools risk exacerbating rather than alleviating disparities in mental health care. In parallel, Elyoseph et al. (2024) frame these technical and clinical risks within a broader ethical discussion about the democratization of mental health care through generative AI. They argue that while GenAI tools offer the potential to increase access, personalization, and theoretical flexibility in mental health support, they also risk reinforcing existing power hierarchies, deepening epistemic dependence on corporate-controlled systems, and creating misleading perceptions of therapeutic authority. The authors propose a strategic questionnaire for evaluating GenAI mental health tools, emphasizing transparency, cultural responsiveness, and safeguards against overreliance. Their sociohistorical and philosophical framing underscores that responsible design must go beyond functional improvements to include ethical commitments to equity, decentralization, and user empowerment.

To address these challenges, scholars advocate for participatory design models that actively involve affected communities throughout the AI development lifecycle. Such models ensure that the lived experiences of marginalized populations inform the creation and implementation of AI tools, leading to more equitable and effective outcomes. Yet, as Delgado et al. (2023) point out, current practices in participatory AI often fall short of this vision. “AI researchers and practitioners are adopting tactics… that may reify or amplify existing power dynamics,” such as relying on proxies or limited input during narrow parts of the design cycle. Their analysis reveals that despite good intentions, much participatory work remains consultative rather than collaborative or co-creative, undermining the potential for genuine stakeholder agency. By involving users not just as testers but as co-designers, developers can uncover blind spots that might otherwise go unnoticed such as culturally specific expressions of distress or trust-related barriers to disclosure. This approach also fosters greater transparency and accountability, as the communities most impacted by these technologies play a role

in shaping their direction (Borning, Friedman, & Kahn, 2004). Moreover, participatory design helps build trust and long-term engagement, which are essential for mental health interventions where sustained use and user comfort are key to efficacy. In centering community knowledge, participatory frameworks move beyond a one-size-fits-all mentality and toward genuinely inclusive innovation.

A high-profile example of these concerns materialized in 2023, when the Federal Trade Commission (FTC) charged BetterHelp—a popular online therapy platform—with violating user trust by sharing sensitive mental health data with advertisers such as Facebook and Snapchat, despite promising confidentiality. The FTC's investigation revealed that BetterHelp used consumers' email addresses, IP addresses, and answers to intake questionnaires for marketing purposes without informed consent. As a result, the company agreed to a $7.8 million settlement and was banned from sharing such data in the future (FTC, 2024). The BetterHelp case illustrates the dangers of insufficient privacy protections in AI-driven mental health services and underscores that therapeutic chatbots, even when widely used, may fail to uphold the ethical standards expected in clinical care. This incident has amplified calls for stronger regulation, transparency, and participatory design that centers the experiences and consent of vulnerable users. It also reinforces concerns that mental health chatbots must be held to standards equivalent to those in clinical practice, especially as they scale into sensitive domains of human emotion and psychological well-being. While corporate misuse of data underscores systemic risks, user perspectives further reveal how these issues manifest on the ground.

Student Voices: An Interview Study on AI Mental Health Chatbot Use in Marginalized Communities

Method

To explore how AI mental health chatbots are experienced by underserved students in real-world contexts, I conducted a qualitative interview study with 18 community college students who identified as low-income, first-generation college students of color. This demographic has been largely underrepresented in prior chatbot evaluations, which have typically drawn on general or self-selected samples with limited attention to race, class, or educational background (Gaffney et al., 2019; Yuan et al., 2025). Participants were recruited through student equity centers, wellness programs, and cultural clubs across three urban community colleges. All participants provided informed consent through a process that emphasized voluntary participation, anonymity, and the ability to withdraw at any time. The study was reviewed and approved by a university-affiliated Institutional Review Board (IRB), with special attention to the emotional risks and potential benefits of asking participants to discuss their mental health and digital help-seeking experiences.

Students were asked to engage with a publicly available mental health chatbot of

their choosing for at least 10 minutes a day over a 10-day period. Rather than prescribing specific chatbot tools, the study used an open-choice approach that allowed students to engage with whatever platforms felt most relevant or accessible to them. This real-world design emphasized autonomy, reflected naturalistic usage patterns, and reduced potential bias linked to specific platform reputations or interface differences. As such, the identities of the tools used are not disclosed in the study's findings. Following the usage period, participants took part in a semi-structured, 30–45 minute interview conducted via Zoom or in person. Interviews were designed to elicit open-ended reflections on their experiences with the chatbot, including emotional resonance, trustworthiness, cultural fit, and potential for long-term use. Sample questions included:

- "What was your first impression of the chatbot?"
- "Did you feel like it understood you?"
- "Would you turn to it again if you were struggling?"
- "What felt missing or uncomfortable?"

All interviews were transcribed and coded using NVivo 14. A hybrid coding approach was employed: deductive codes were drawn from existing literature (e.g., privacy, empathy, accessibility), while inductive codes were developed through repeated close reading of transcripts to surface themes specific to the experiences of students of color navigating mental health stigma, digital tools, and trust. A second coder was trained to review a subset of transcripts, and discrepancies were discussed to improve inter-coder reliability.

This study contributes novel empirical data to a literature that has so far relied heavily on surveys, clinical trials, and general-population evaluations. Most importantly, it centers voices that are too often treated as afterthoughts in the design and evaluation of mental health technologies. By focusing on marginalized students who are both disproportionately affected by mental health disparities and least likely to access traditional therapy, this study offers grounded insights into the real-world barriers and emotional logics shaping chatbot use among those most in need—and most at risk of being left behind by digital health solutions. My small study parallels a Stanford-based project led by Dr. Harriett Jernigan, who developed the first culturally tailored mental health chatbot for Black students using a protected, ethically sourced dataset built from 100 hours of interviews and 1,000 question-and-answer pairs derived from interviews between six Black therapists and six Black students. Preliminary results from that project, shared with me during the research process, showed markedly higher trust, user satisfaction, and culturally resonant guidance compared to commercial chatbots. The comparison highlights the transformative potential of culturally grounded, community-driven design and further contextualizes the critiques voiced by students in my study.

Findings

Despite selecting different chatbots, students consistently echoed three core

themes: lack of cultural attunement, shallow emotional connection, and concerns about long-term trust. Many found the chatbot interfaces friendly and appreciated having access at any time, but ultimately felt that the tools failed to capture the complexity of their lived experiences. Several participants described a sense of being "talked at" rather than truly heard, citing “cringe,” “bland,” “dismissive,” repetitive responses that always tried to “find some excuse for racism,” “asked for more context,” “asked if I was being overly sensitive” or offered vague affirmations that felt generic rather than personalized. While some students noted moments of temporary relief or distraction, most said they would not continue using the chatbot regularly. Their critiques centered less on specific technical features and more on the bots’ inability to understand racialized stress, cultural stigma, or intersectional forms of marginalization. As one participant put it, “It’s like it wants me to journal, but it doesn’t know what I’ve been through or how I talk about stress in my family.” The majority of students said chatbots were a “good place to start” and appreciated their privacy and round-the-clock availability. Many respondents reported feeling emotionally safe in the initial stages of using a chatbot, especially when they were unsure about speaking to an adult. However, trust remained a barrier: most students said they would not rely on chatbots regularly, expressing concerns about authenticity, effectiveness, and whether the bots could truly understand their experiences. A recurring theme was the desire for mental health tools developed *by and for* their own communities—tools that reflect their cultural backgrounds, lived realities, and communication styles. These findings highlight a crucial gap: while AI tools can increase access, they must also earn trust through cultural specificity and community involvement. Importantly, students were not anti-chatbot; rather, they wanted AI tools to feel relevant, trustworthy, and human-informed. Across all interviews, there was strong consensus that any digital mental health intervention must be built in collaboration with the communities it aims to serve, incorporating culturally specific understandings of care, resilience, and communication. When viewed through the lens of equity and inclusion, student feedback offers valuable guidance for developers and policymakers seeking to create more effective, ethical, and engaging chatbot interventions.

Discussion

Despite their accessibility, chatbots fall short in several critical areas. First and foremost is the question of empathy. While AI can mirror therapeutic language, studies comparing ChatGPT-like systems with licensed counselors show a stark difference in emotional depth and connection (Eryılmaz & Başal, 2024). More importantly, chatbots remain unable to handle mental health crises. In moments of suicidal ideation or psychosis, AI may offer inappropriate or generic advice—or fail to escalate at all. While some platforms now direct users to suicide hotlines, this remains inadequate for real-time intervention.

Ethical and legal risks also abound. AI systems can propagate biased, exclusionary, or culturally insensitive advice due to limitations in their training data. For example, Tawiah and Monestime (2024) highlight how language-based AI models often fail non-English-speaking immigrants, adolescents, people with disabilities, and rural residents due to a lack of culturally and contextually relevant training data. This failure leads to exclusionary or inappropriate advice and can reinforce longstanding disparities in access and outcomes. Without explicit efforts to develop linguistically and culturally inclusive systems, these communities may be further marginalized by the very tools intended to support them. Furthermore, privacy violations have already occurred. The FTC fined BetterHelp in 2023 for illegally sharing sensitive user data with advertisers, raising fears about surveillance and commodification of mental illness (Federal Trade Commission, 2024).

Regulatory Gaps and Participatory Injustice

One of the most pressing concerns in the mental health chatbot ecosystem is the lack of a robust regulatory framework. Human therapists operate under clearly defined legal and ethical obligations, including licensure, confidentiality rules, and professional standards of care. By contrast, chatbot developers currently face far fewer accountability mechanisms. When a chatbot fails to respond to a crisis, offers inappropriate guidance, or mishandles sensitive data, there is limited recourse for users and often no clear party responsible for harm (Pozzi & De Proost, 2024). Equally troubling is the ongoing risk of participatory injustice.

To address these gaps, I propose a multi-pronged policy framework:

- **Federal and State Guidelines for Clinical AI:** Establish clear regulatory standards under agencies like the FDA or HHS for digital mental health interventions, including chatbot use in therapeutic contexts. These standards should mirror protections in place for traditional mental health care, ensuring that AI tools do not bypass clinical responsibility.
- **Certification and Auditing Bodies:** Create independent certification boards to regularly evaluate mental health chatbots on criteria such as cultural competence, crisis response accuracy, privacy safeguards, and bias mitigation. These bodies can operate akin to educational or health accreditation organizations.
- **Community Advisory Panels:** Require companies to form standing advisory panels composed of mental health professionals and community members from underrepresented groups. These panels should have input in both product development and periodic review.
- **Data Transparency and Consent Reform:** Mandate transparent data collection, storage, and usage policies. Users should be able to see how their data is being used, give granular consent, and opt out without

penalty. Companies should be penalized for violations, including deceptive data sharing with advertisers.

- **Public Funding for Culturally Specific Innovation:** Allocate federal or philanthropic funding toward the creation and support of non-commercial, open-source chatbot platforms developed by and for marginalized communities. Examples such as Dr. Harriet Jernigan's Stanford project should be studied and scaled.
- **Integrated Human-AI Care Models:** Encourage health systems and institutions to pair chatbot use with trained human oversight, especially when AI tools are deployed in high-risk settings like colleges, workplaces, or health clinics.

Through this framework, policy can play an active role not in stifling innovation but in channeling it toward inclusive, accountable, and clinically sound applications. Ethical deployment of mental health chatbots depends not only on good intentions but on structural protections and active community participation.

Conclusion

This study explicitly avoids comparing or ranking specific commercial chatbot tools. Instead, it reveals a common pattern: despite choosing different AI mental health tools, students from low-income, first-generation communities of color shared strikingly similar concerns around cultural fit, emotional resonance, and trust. These shared experiences suggest that current commercial chatbots, while useful as a first step, have not yet achieved the inclusivity or contextual depth needed to fully meet the needs of these populations.In contrast, preliminary findings from Dr. Harriett Jernigan's Stanford study, featuring a chatbot trained on a protected dataset of dialogues between Black therapists and Black students, demonstrates the potential of culturally responsive design. Students in that study showed greater engagement and trust, indicating that mental health tools grounded in the language, experiences, and values of specific communities can dramatically shift user outcomes.

Thus, this paper calls for an expanded vision: one that incorporates participatory design, robust data protection, and culturally specific guidance. Developers, funders, and regulators must recognize that digital mental health tools will only reach their full potential when they are built with and not just for the communities they serve. By integrating empirical evidence, lived experience, and interdisciplinary scholarship, this research urges a shift in the conversation: from convenience to care, from generalization to specificity, and from accessibility alone to belonging and trust. In doing so, we can ensure that AI mental health tools evolve into ethical, equitable, and effective bridges to long-term well-being. Future work should continue to amplify community-led design models and rigorously evaluate chatbot performance in diverse, real-world

settings—ensuring that mental health technologies truly evolve to serve those who need them most

References

Abd-Alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., & Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. *International journal of medical informatics*, *132*, 103978.

Aljamaan, F., Temsah, M. H., Altamimi, I., Al-Eyadhy, A., Jamal, A., Alhasan, K., ... & Malki, K. H. (2024). Reference hallucination score for medical artificial intelligence chatbots: development and usability study. *JMIR Medical Informatics*, *12*(1), e54345.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

Birhane, A. (2021). Algorithmic injustice: a relational ethics approach. *Patterns*, *2*(2).

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Borning, A., Friedman, B., & Kahn Jr, P. H. (2004, January). Designing for human values in a urban simulation system: Value sensitive design and participatory design. In *PDC* (pp. 68-71).

Casu, M., Triscari, S., Battiato, S., Guarnera, L., & Caponnetto, P. (2024). AI chatbots for mental health: A scoping review of effectiveness, feasibility, and applications. *Applied Sciences, 14*(13), 5889. https://doi.org/10.3390/app14135889

Delgado, F., Yang, S., Madaio, M., & Yang, Q. (2023, October). The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-23).

Elyoseph, Z., Gur, T., Haber, Y., Simon, T., Angert, T., Navon, Y., ... & Asman, O. (2024). An ethical perspective on the democratization of mental health with generative AI. *JMIR Mental Health*, *11*, e58011.

Eryılmaz, A., & Başal, A. (2024). Rational AIs with emotional deficits: ChatGPT vs. counselors in providing emotional reflections. *Current Psychology, 43*, 34962–34977. https://doi.org/10.1007/s12144-024-06947-w

Federal Trade Commission. (2024, May). BetterHelp customers will begin receiving notices about refunds related to 2023 privacy settlement. https://www.ftc.gov/news-events/news/press-releases/2024/05/betterhelp-customers-will-begin-receiving-notices-about-refunds-related-2023-privacy-settlement-ftc

Fleming, T., Bavin, L., Lucassen, M., Stasiak, K., Hopkins, S., & Merry, S. (2018). Beyond the trial: systematic review of real-world uptake and engagement with digital self-help interventions for depression, low mood, or anxiety. *Journal of medical Internet research*, *20*(6), e199.

Ho, A., Hancock, J., & Miner, A. S. (2018). Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *Journal of Communication*, *68*(4), 712-733.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1-55.

Liang, W., Tadesse, G. A., Ho, D., Fei-Fei, L., Zaharia, M., Zhang, C., & Zou, J. (2022). Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence*, *4*(8), 669-677.

Limpanopparat, S., Gibson, E., & Harris, A. (2024). User engagement, attitudes, and the effectiveness of chatbots as a mental health intervention: A systematic review. *Computers in Human Behavior: Artificial Humans*, 100081.

Marko, J. G. O., Neagu, C. D., & Anand, P. B. (2025). Examining inclusivity: The use of AI and diverse populations in health and social care—a systematic review. *BMC Medical Informatics and Decision Making, 25*, 57. https://doi.org/10.1186/s12911-025-02884-1

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, *2*(11).

Pozzi, G., & De Proost, M. (2024). Keeping an AI on the mental health of vulnerable populations: Reflections on the potential for participatory injustice. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00523-5

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*.

Revathi, K., Priyanka, S., Priyadharshini, S., Niranjan, N. N., Visalachi, N., & Sumathi, P. (2025, January). AI-Driven Approaches to Enhancing Mental

Wellbeing and Stress Relief. In *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)* (pp. 925-931). IEEE.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature medicine*, *29*(8), 1930-1940.

Tawiah, N., & Monestime, J. P. (2024). Promoting equity in AI-driven mental health care for marginalized populations. *Proceedings of the AAAI Symposium Series, 4*(1), 323–327. https://doi.org/10.1609/aaaiss.v4i1.31810

Yuan, A., Garcia Colato, E., Pescosolido, B., Song, H., & Samtani, S. (2025). Improving workplace well-being in modern organizations: A review of large language model-based mental health chatbots. *ACM Transactions on Management Information Systems*, *16*(1), 1-26.