Ethical Implications of Mental Health Chatbots: Addressing Anthropomorphism, Deception, and Regulatory Gaps

THOMAS LEIS, Leibniz Universität Hannover, Hannover, Germany

Mental Health Chatbots (MHCBs) present both opportunities and ethical concerns in mental healthcare. While they promise greater accessibility and immediate support, their design and marketing may foster anthropomorphic perceptions that may lead to user over-reliance and potential deception. Current debates oscillate between optimism, emphasizing MHCBs' potential to fill gaps in mental healthcare, and skepticism, highlighting risks such as misinformation and inadequate regulatory oversight. However, both perspectives often lack a structured ethical foundation. This paper sketches a novel ethical framework grounded in principles entrenched in bioethics and AI Ethics to establish clear guidelines for responsible MHCB design and deployment. By shifting the focus from whether MHCBs should be used to how they should be ethically implemented, this framework provides concrete recommendations for policymakers, developers, deployers, and professional therapists to ensure MHCBs align with ethical and clinical standards.

Keywords: Mental Health Chatbots, Marketing, Anthropomorphism, Deception, Human-AI Interaction, Bioethics, Regulation of AI, Responsible AI Design and Deployment

Reference Format:
Thomas Leis. 2025. Ethical Considerations in Mental Health Chatbots: Addressing Anthropomorphism, Deception, and Regulatory Gaps. In *Proceedings of the Fourth European Workshop on Algorithmic Fairness (EWAF'25)*. Proceedings of Machine Learning Research, 7 pages.

1 Introduction

In recent years, the global mental health crisis has worsened. Barriers to seeking therapy include shortages in available services, lack of symptom recognition, high treatment costs, lack of time, and concerns about confidentiality and stigma [1]. Against this backdrop, mental health chatbots (MHCBs) have emerged as a potential tool to bridge accessibility gaps and alleviate the scarcity of professional services.

This paper argues that MHCBs raise novel ethical concerns, particularly regarding anthropomorphism and deception. Developers may design their systems to mimic human therapists by incorporating empathetic language, conversational fluency, and emotional responsiveness [2]. Combined with users' natural tendency to ascribe human-like understanding to AI [3], this may lead to misplaced trust and over-reliance on systems. While human therapists are guided by professional training and ethical codes [4], current MHCBs are classified as general wellness products [5], exempting them from the strict ethical and safety requirements applied to medical devices [6].

Authors' Contact Information: Thomas Leis, Leibniz Universität Hannover, Hannover, Germany, thomas.leis@stud.uni-hannover.de.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

The existing literature reflects a polarized debate: optimists emphasize the potential of MHCBs to expand access, reduce clinician workload, and provide immediate support [7-9]. Skeptics point to risks such as misinformation, the loss of human empathy in care, users overrelying on MHCBs, and inadequate regulation [10-12]. However, both perspectives rarely address the core bioethical principles of autonomy, beneficence, non-maleficence, and justice in their respective analyses [13]. These principles provide a structured approach to assess the ethical implications of MHCBs. Autonomy concerns whether users can make informed decisions when interacting with MHCBs; beneficence requires MHCBs to genuinely support user well-being; non-maleficence demands proactive harm prevention; and justice raises concerns about equal access and potential AI-driven biases.

The current debate centers on *whether* MHCBs should play a role in mental healthcare rather than *how* they should be responsibly designed and deployed in alignment with clinical, ethical, and regulatory frameworks. To address this gap, I sketch an ethical framework that synthesizes (1) principles entrenched in bioethics and AI Ethics, and (2) connect these principles to stakeholder groups most involved with MHCBs. This framework helps move beyond whether MHCBs should be built to focus on how they may be designed and deployed to maximize benefits while minimizing risks.

2 Current Efforts to Build MHCBs

MHCBs are increasingly being used to support people struggling with mental health issues [14]. Today's MHCBs include smartphone apps such as Wysa, and Woebot Health and others, which typically offer: (a) text-based chats and (b) guided self-help activities, usually based on cognitive behavioral therapy (CBT) and other mindfulness techniques [15]. The chatbot asks users about their emotions and how they are feeling and suggests one of their self-help activities, such as meditation, or breathing exercises, based on user responses.

In terms of how MHCBs arrive at their responses, Woebot and Wysa work with rule-based, pre-written scripts that are produced by in-house writing teams and reviewed and approved by professional therapists [16, 17]. Providers are also required to state that their MHCBs, in the case of Woebot, "have not been evaluated, cleared, or approved by the FDA" [18] and are therefore unsuitable for emergencies, and urge users to seek professional help.

However, such disclaimers are at odds with prominent marketing claims that position MHCBs as alternatives when traditional therapy is not accessible. Wysa, for example, advertises being "available 24/7" and "utilizing evidence-based tools like CBT to provide clinically effective support" [19]. Woebot Health states, "A severe lack of therapists, confusing insurance jargon, and scheduling headaches – the mental health care system is unable to support all who need it" [20] and positions itself as a solution to the current mental health care system, which they believe is "unable to support everyone who needs it".

Such claims raise ethical concerns, as mental health care is a sensitive area involving vulnerable patients, power imbalances between patients and therapists, and risks of harm due to inadequate treatments. While professional therapists undergo years of training and education and adhere to strict ethical codes [4], MHCBs operate within a regulatory gray area, often classified as "general wellness products" [5] rather than medical devices, thus exempting them from stricter oversight [6]. This regulatory gap, combined with marketing strategies like the above, creates a risk that users may develop undue trust in chatbot capabilities, believing them to be equivalent to, or better than, human professional therapy.

3 The Current Debate: Optimistic and Skeptical Perspectives on MHCBs

Optimists argue that the increased accessibility of chatbots improves mental health care. For example, [7] and [8] claim that 24/7 access to chatbots reduces barriers to care, but do not consider whether this access leads to meaningful user autonomy and informed decision making. In addition, over-reliance on chatbots may create a false sense of security and delay necessary human intervention.

Other optimists claim that MHCBs may simulate empathy consistently, without being influenced by factors that can affect human therapists, such as implicit bias (e.g., based on a user's race, gender, or group affiliation) or fluctuating emotional states caused by stress, fatigue, or personal circumstances [9]. From this perspective, MHCBs are seen as offering a more neutral support environment. However, when MHCBs prioritize superficial emotional validation over deeper therapeutic engagement, they risk creating emotionally hollow experiences that leave users with a false perception of emotional support, raising questions about what kind of support they truly provide.

Conversely, skeptics highlight concerns such as therapeutic misconceptions, referring to users who "underestimate the restriction of such technologies and overestimate their ability to provide therapeutic support and guidance" [10, p. 2], which may lead users to delay seeking professional help. Other critics point to a lack of genuine human interaction [11], and questions about whether narrow AI systems such as chatbots can deliver therapy effectively [12]. Yet, these critiques are rarely accompanied by suggestions for how MHCBs might be designed and deployed more responsibly.

Ultimately, both optimistic and skeptical perspectives fall short of providing a normative foundation for evaluating MHCBs ethically. Without a structured ethical framework that translates entrenched bioethical principles, professional codes, and regulatory standards into concrete design and deployment criteria, the debate remains fragmented rather than constructively guiding responsible development and deployment.

To address this gap, I sketch an ethical framework for the responsible design and deployment of MHCBs. Integrating debates on the ethical nature of mental healthcare, the current regulatory landscape of MHCBs, and analyzing anthropomorphism and deception as key ethical concerns in human-AI interactions, my framework aims to provide structured guidance for policymakers, developers, and mental health professionals, ensuring that MHCBs align with moral principles while mitigating identified risks.

4 Anthropomorphism and Deception in AI Chatbots

A key factor shaping user interaction with MHCBs is *anthropomorphism*, the tendency to attribute human-like properties, characteristics, or mental states to non-human entities [3]. Developers may anthropomorphize these systems *by design* with human-like voices, specific dialogue styles, visual cues, and avatars to further enrich the chatbot's human-like communication, thus enhancing user acceptance, engagement, and retention [2]. Moreover, users may also independently project human-like attributes onto chatbots, even without developer intention (see [21] for a historical precursor). While some argue that anthropomorphizing chatbots can improve the overall quality of human-AI interactions by making these systems more familiar and effective [22], critics warn that such design decisions and human-AI interactions can lead users to develop false expectations of MHCBs and misapply characteristics normally associated with trained and licensed human therapists, such as genuine emotional understanding, empathy, and therapeutic knowledge and competence. The author of [23] argues that such misattributions and misplaced trust in conversational AI systems due to what she calls *anthropomorphic false beliefs* or "a mistaken belief that an entity possesses human-like traits when, in fact, it does not" [23, p. 1] This is especially relevant in mental health contexts and thus for MHCBs, as anthropomorphism may both encourage users to willingly self-disclose sensitive personal information to MHCBs due to perceived user anonymity, trust in MHCBs, and less fear of judgment [24], thus raising ethical concerns surrounding the privacy and data protection of vulnerable users. This leads to the question of when anthropomorphism would be permissible in the context of MHCBs. Unfortunately, a clear consensus does not yet exist because the downstream effects of anthropomorphism are largely dependent on user perception, which also begs the question of how much responsibility designers and deployers can bear for effects that are, at least in part, beyond their control [25].

5 An Ethical Framework for Responsible Design and Deployment of MHCBs

An ethical framework for MHCBs is needed to address concerns about anthropomorphism and deception, regulatory gaps, the potential for overreliance by users due to misleading marketing and design decisions, and ethical uncertainties about their role in mental healthcare. Here, I outline a framework that draws on and integrates nine established values from the bioethics and AI ethics literature. The first four values, namely ***autonomy***, ***beneficence***, ***non-maleficence*** and ***justice***, derive from Beauchamp and Childress' four *Principles of Biomedical Ethics* [13], which guide ethical decision-making in healthcare, especially in a context involving vulnerability, trust, and informed consent. The remaining five values, which are ***transparency***, ***accountability***, ***human oversight***, ***trust calibration***, and ***privacy and data protection*** are adapted from leading AI ethics frameworks, namely the IEEE Ethically Aligned Design guidelines [26] and the European Commission's Ethics Guidelines for Trustworthy AI [27]. Additionally, emerging legal frameworks, such as the EU AI Act [28], begin to translate ethical concerns, particularly transparency, accountability, human oversight, and data governance, into legally binding obligations.

Based on these core ethical values, my framework allows deriving concrete principles for MHCB development and deployment: 1) MHCBs must uphold *autonomy* by enabling informed user decisions, while 2) *beneficence* ensures they provide genuine support without harm. 3) *Nonmaleficence* demands proactive safeguards against potential harms such as misinformation stemming from false and misleading chatbot responses. 4) *Justice* requires equitable access, regardless of socioeconomic status. For example, developers/deployers should design MHCBs to be economically accessible to different populations by minimizing costs, supporting multiple languages, and ensuring usability for different levels of users' digital literacy. 5) *Transparency* necessitates clear disclosure of the chatbot's precise role, capabilities, and limitations to prevent users overestimating the system's capabilities. 6) *Accountability* must clarify the partial responsibilities of developers and deployers in design and marketing. 7) *Human oversight* is crucial, especially in crises requiring professional intervention. 8) *Trust calibration* aims for an appropriate level of user trust when using chatbots that is not too low to cause skepticism and not too high to cause undue trust and overconfidence.

Finally, 9) *privacy and data protection* requires clear communication to users about how their personal data is stored and used.

Further developing this framework also involves identifying the key stakeholders who will interact with MHCBs. I distinguish three core groups: 1) users, 2) professional therapists, and 3) developers/deployers. Users are the primary beneficiaries, but also the most vulnerable to the consequences of ethical or unethical MHCB design. Therefore, *transparency* is essential to prevent users from misjudging the role, capabilities, and limitations of MHCBs, thereby enabling autonomous decision-making about their use. Therapists may benefit from a reduced workload, but also risk being sidelined if users (unduly) prefer MHCBs over human-led therapy. *Human oversight* is therefore necessary to ensure that MHCBs are a complement, not a replacement, for therapy. AI developers/deployers significantly shape human-AI interaction through design and marketing. Thus, *transparency* is crucial to avoid misleading or exaggerated claims, contributing to *beneficence* by helping users rather than harming them. It also supports *nonmaleficence*, by preventing misinformation or ambiguities that could cause harm to users. Finally, *privacy and data protection* concern developers and deployers, in regard to informing users about how user' data are stored ans used, and it also concerns users themselves, in regard to making informed choices about how to interact with MHCBs.

6 Conclusion

MHCBs present both opportunities and challenges for mental healthcare. While they may have the potential to provide more accessible support, their ethical risks, including anthropomorphism and deception, remain underrecognized. Regulatory gaps and uncertainties surrounding MHCBs' status as either a general wellness device or approved medical device necessitate further clarification. The academic debate between optimists and skeptics lacks a structured ethical grounding, leaving fundamental questions about responsible design and deployment unanswered. To address this, this paper has sketched a novel ethical framework that synthesizes key bioethical principles, AI Ethics insights and their relation to key stakeholder groups, to ensure MHCBs are developed and deployed responsibly.

Acknowledgements

The author thanks Dr. Donal Khosrowi for his contributions in supervising this project and shaping and refining the ideas developed here.

References

[1] Şerife Tekin. 2021. Is big data the new stethoscope? Perils of digital phenotyping to address mental illness. *Philosophy & Technology*, *34*(3), 447-461. https://doi.org/10.1007/s13347-020-00395-7

[2] Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023. Mirages: On anthropomorphism in dialogue systems. *arXiv preprint arXiv:2305.09800*, 1-15.

[3] Nicholas Epley, Adam Waytz, and John T. Cacioppo. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, *114*(4), 864-886. https://doi.org/10.1037/0033-295X.114.4.864

[4] American Psychological Association. 2017. *Ethical principles of psychologists and code of conduct* (2002, amended effective June 1, 2010, and January 1, 2017). https://www.apa.org/ethics/code/

[5] Center for Devices and Radiological Health, US Food and Drug Administration. General Wellness. Policy for low risk devices. *FDA*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/general-wellness-policy-low-risk-devices. Accessed: March 12, 2025.

[6] Julian De Freitas and I. Glenn Cohen 2024. The health risks of generative AI-based wellness apps. *Nature medicine*, *30*(5), 1269-1275. doi: https://doi.org/10.1038/s41591-024-02943-6

[7] Olivia Nelson. 2023. Revolutionizing Mental Health Care: The Impact of AI Chatbots. *Infotech Journal Scientific and Academic*, *4*(1), 197-218.

[8] Rocco de Filippis and Abdullah Al Foysal. 2024. Chatbots in psychology: Revolutionizing clinical support and mental health care. *Voice of the Publisher*, *10*(3), 298-321. doi: 10.4236/vp.2024.103025

[9] Michael Inzlicht, C. Daryl Cameron, Jason D'Cruz and Paul Bloom. 2024. In praise of empathic AI. *Trends in Cognitive Sciences*, *28*(2), 89-91.

[10] Zoha Khawaja and Jean-Christophe Bélisle-Pipon. 2023. Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots. *Frontiers in Digital Health*, *5*, 1278186, 1-13. doi: 10.3389/fdgth.2023.1278186

[11] Julia E. H. Brown and Jodi Halpern. 2021. AI chatbots cannot replace human interactions in the pursuit of more inclusive mental healthcare. *SSM-Mental Health*, *1*, 100017, 1-5. doi: 10.1016/j.ssmmh.2021.100017

[12] J. P. Grodniewicz and Mateusz Hohol. 2023. Waiting for a digital therapist: three challenges on the path to psychotherapy delivered by artificial intelligence. *Frontiers in Psychiatry*, *14*, 1190084, 1-12. doi: 10.3389/fpsyt.2023.1190084

[13] Basil Varkey. 2021. Principles of clinical ethics and their application to practice. *Medical Principles and Practice*, *30*(1), 17-28. https://doi.org/10.1159/000509119

[14] Luke Balcombe. 2023. AI chatbots in digital mental health. In *Informatics* (Vol. 10, No. 4, p. 82). MDPI, 1-16. https://doi.org/10.3390/informatics10040082

[15] Eoin Fullam. 2025. How does a mental health chatbot work? A 'conversation design'concept of mental health intervention. *History of the Human Sciences*, 09526951241305274, 1-23. https://doi.org/10.1177/09526951241305274

[16] Alison Darcy. 2023. *Why Generative AI Is Not Yet Ready for Mental Healthcare.* Woebot Health. https://woebothealth.com/why-generative-ai-is-not-yet-ready-for-mental-healthcare/. Accessed: March 11, 2025.

[17] Christel Lynne Chang, Chaitali Sinha, Madhavi Roy, and John Chee Meng Wong. 2024. AI-Led Mental Health Support (Wysa) for Health Care Workers During Covid-19: Service evaluation. *JMIR Formative Research*, *8*, e51858. doi:10.2196/51858

[18] Woebot for Adults. Woebot Health. https://woebothealth.com/instructions-for-use/. Accessed: March 12, 2025.

[19] Wysa. 2024. *Wysa—Mental Health Support*. Wysa. https://www.wysa.com/. Accessed: March 10, 2025.
[20] Woebot Health. 2024. *Woebot—The Mental Health Ally*. Woebot Health. https://woebothealth.com/. Accessed: March 10, 2025.

[21] Joseph Weizenbaum. 1976. *Computer power and human reason: From judgment to calculation.* W. H. Freeman & Co.

[22] Luisa Damiano and Paul Dumouchel. 2018. Anthropomorphism in human–robot co-evolution. *Frontiers in psychology*, *9*, 468, 1-9. https://doi.org/10.3389/fpsyg.2018.00468

[23] Beatrice Marchegiani. 2025. Anthropomorphism, False Beliefs, and Conversational AIs: How Chatbots Undermine Users' Autonomy. *Journal of Applied Philosophy*, 1-18. https://doi.org/10.1111/japp.70008

[24] Marita Bjaaland Skjuve and Petter Bae Brandtzæg. 2018. Chatbots as a new users interface for providing health information to young people. In: *Youth and news in a digital media environment–Nordic-Baltic perspectives*, 59-66.

[25] Canfer Akbulut, Laura Weidinger, Arianna Manzini, Iason Gabriel, and Verena Wieser. 2024. All Too Human? Mapping and Mitigating Risks from Anthropomorphic AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Vol. 7, 13-26. https://doi.org/10.1609/aies.v7i1.31613

[26] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2017. *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems* (Version 2). IEEE. Retrieved from http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

[27] European Commission, High-Level Expert Group on Artificial Intelligence. 2019. Ethics Guidelines for Trustworthy AI. Brussels: European Commission. Retrieved from: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[28] European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. Retrieved from: https://eur-lex.europa.eu/eli/reg/2024/1689/oj