The Promise and Pitfalls of Mental Health AI Chatbots

Riley Little, Navya Varda, Shayna Patel, Medha Nagaluri, William Banks Leavitt

DATA 120
Ethics of AI Spring 2025

Question and Background

Our **research question** is: What are the ethical advantages and risks of using AI-powered mental health chatbots?

These chatbots use artificial intelligence to provide mental health support through automated, text-based conversations and became prominent during the COVID-19 pandemic due to the mental effects of isolation and limited access to care.

An example is **Therabot**, a popular mental health AI chatbot that uses natural language processing (NLP) to simulate therapeutic dialogue and is trained on therapist-patient Cognitive Behavioral Therapy (CBT).

Arguments

Pros and Cons of Mental Health AI Chatbots

Pros	Cons
Increases access to mental health support	Lacks genuine human empathy
Available 24/7	Privacy and data security risks
Low-cost and anonymous	Potential for misinformation
Reduces stigma around seeking help	Inadequate response to crisis situations

1. **Accessibility** and **affordability** are key benefits, but cannot replace genuine human empathy.
2. **Anonymity** and **24/7 support** reduce stigma, yet raise risks of misinformation and poor crisis response.
3. **Fairness** depends on quality training data. Biased or incomplete data may harm vulnerable users.
4. **Privacy** and **security protections** are essential to avoid exposing sensitive mental health information.
5. Ethical use requires **balance**: chatbots should support, not replace, human care.

Policy Recommendations

1. **Supervised Use Only for Crisis Response:** AI mental health chatbots must operate under the supervision of a licensed therapist who can monitor flagged conversations and step in when needed to ensure adequate crisis response.

2. **Mandatory Transparency:** Users must be clearly informed that they are interacting with an AI, understand the lack of real human empathy, and know how their data is being used.

3. **Bias and Fairness Audits:** Chatbots must participate in regular audits to detect and correct for bias to ensure fair and equitable support across diverse user groups.

4. **Certified Evidence-Based Training:** All chatbot content must be based on peer-reviewed, evidence-based therapeutic practices (e.g., CBT), comply with State laws, and reviewed by the National Institute of Mental Health, a technical board, and World Health Organization to ensure no unethical advice is given.

Course Integration and Literature

Philosophical:

- Our project uses a **utilitarian** ethical framework, focusing on outcomes and maximizing overall well-being
- For mental health AI chatbots, this means improving access while weighing risks like privacy issues, misinformation, and poor crisis response.

Legal:

- **Privacy protections**, **disclosure** that users are interacting with AI, and **vetting of training data** by medical professionals are essential.
- **Escalation protocols** must ensure users in crisis are directed to human help.

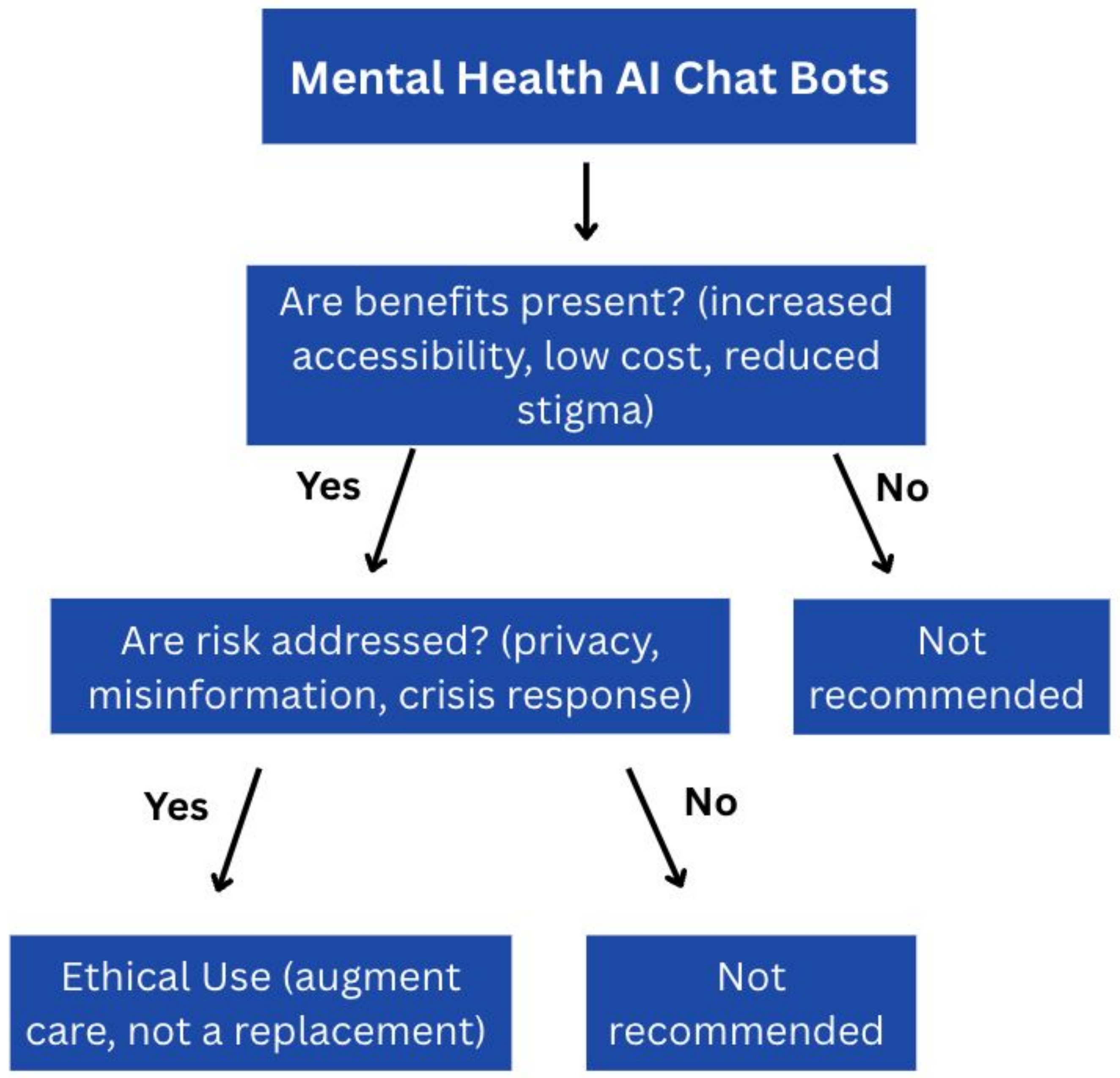
Technological:

- Chatbot design should prioritize **safety**, **reliability**, **transparency,** and **accessibility.**
- Systems should be **audited** by mental health professionals and trained on evidence-based conversations.
- Dropout rates can be improved with better **UI**, **personalization**, and **feedback**.

Synthesis:

- Mental health chatbots increase **accessibility** and **affordability** but pose serious **risks**.
- Maximizing well-being requires strong monitoring, ethical data practices, and human oversight to balance innovation with user safety.

Results



Mental Health AI Chat Bots

Are benefits present? (increased accessibility, low cost, reduced stigma)

Yes

No

Are risk addressed? (privacy, misinformation, crisis response)

Not recommended

Yes

No

Ethical Use (augment care, not a replacement)

Not recommended

Conclusion

From a utilitarian perspective, AI mental health chatbots offer **benefits** but also come with potential **harms**.

We propose **four key policies** to maximize societal well-being: supervising chatbot use, requiring transparency, running fairness audits, and certifying evidence-based training data.

Even with these in place, AI chatbots should be used alongside a licensed therapist, **not as a replacement.** As the mental health crisis continues to grow, chatbots can help reduce obstacles to care, but they are only a temporary solution and cannot replace the empathy and judgment of a human therapist.

References:

Kühler, M., Williams, B., & Nida-Rümelin, J. (2024). *Ethical theories*. In M. Kühler, B. Williams, & J. Nida-Rümelin (Eds.), *The Routledge Handbook of Ethics and Artificial Intelligence*. Routledge.

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Dartmouth College. (2025, March 5). *First therapy chatbot trial yields mental health benefits*. Dartmouth News. https://home.dartmouth.edu/news/2025/03/first-therapy-chatbot-trial-yields-mental-health-benefits

Utah State Legislature. (2025). H.B. 452 Artificial Intelligence Amendments. https://le.utah.gov/~2025/bills/static/HB0452.html

Siddals, S., Torous, J., & Coxon, A. (2024). "It happened to be the perfect thing": experiences of generative AI chatbots for mental health. Npj mental health research, 3(1), 48. https://doi.org/10.1038/s44184-024-00097-4

Artificial intelligence in mental health care. (2024, November 21). American Psychological Association. https://www.apa.org/practice/artificial-intelligence-mental-health-care