Ethics Report: Psychological Pre-Consultation Chatbot

Ethical Considerations in Automated Mental Health Support

Why do we need chatbot to help?

In today's fast-paced world, where population growth, economic pressures, and the competitive job market intensifies daily stressors, mental health has become an alarming global concern. According to the World Health Organization (WHO), at least one in eight people worldwide live with a mental disorder of which more than 60% suffer from anxiety or depression (*WHO*, 2025). Despite this growing prevalence, access to mental health support remains limited due to a shortage of trained professionals, long waiting times, high costs, and the stigma still associated with seeking psychological care.

In response to these challenges, technological innovations in artificial intelligence (AI) along with advancement in technologies such as sentiment analysis and NLP are increasingly being explored as a means of delivering accessible, affordable, and personalized mental health support. AI-powered tools such as chatbots, conversational agents, and mobile health applications promise 24/7 availability reducing barriers to care by simulating a dialogue with the users using natural language. CUIs like Woebot, Wysa and Replika are used by a lot of people and have the potential to provide early interventions, emotional support, and even crisis detection for individuals in distress.

However, the integration of AI into such a sensitive domain raises important ethical questions. Issues of privacy, responsibility, bias, and safety must be critically examined to ensure that these systems do not unintentionally cause harm. Saeidnia et al., 2024, reviewed 18 key ethical considerations while utilizing AI to aid mental health of users. This essay will explore some of the ethical considerations of automated mental health support and assess potential risks and outline strategies to mitigate them.

Potential Risks and Mitigation Strategies

Informed Consent and Transparency

One of the major ethical concerns in automated mental health support is the risk of users developing *emotional attachments* to AI chatbots. While these tools can simulate empathy and provide comforting responses, they cannot truly understand or reciprocate human emotions. This becomes problematic when users are unaware that they are interacting with a bot rather than a human. Informed consent and transparency are therefore crucial, and users should fully understand the nature, scope and limitations of AI-based support.

Character AI, which is famous for its generative AI service capable of users creating characters of famous personalities and letting other to chat with them after publishing has been facing with rising concern in recent years. Several news of teens suicide upon developing strong emotional connection with the characters are seen rising in the internet (*Lawsuit Claims Character.AI Is Responsible for Teen's Suicide*, 2024)

On the other hand, Replika AI which was developed to provide a listening ear to the users has this feature to genuinely connect with their users by creating backstories for themselves and talking about their personal life. This has led to situations like increased offline social anxiety, suicidal behaviour, eating disorders, etc (Chow, 2025).

While developing the automated psychological pre-consultation chatbot, the following strategies were used to ensure transparency and informed consent.

Mitigation strategies used:

1. **Disclaimer**

Clear disclaimers and reminders are shown to the user at the start of every conversation and displayed all the time in the sidebar. The disclaimer explicitly conveys its boundaries and limitations.

- It clearly suggests that it is not a licensed therapist and hence cannot provide any medical diagnosis or prescription.
- Things that it can offer like an empathetic and non-judgemental listening ear.
- Suggests users to seek help immediately in situations of crisis.

2. **Promoting self-reflection**

Bot response is moderated and ensured that it does not provide any directed or manipulative commands. It is strictly prohibited from giving medical diagnosis or treatment suggestions. Rather users are encouraged to reflect on their decisions by promoting them to share about their experiences if the input is not categorised as a strong crisis.

To make the model response better, a strict system prompt that enforces role boundaries and limitations is used to get response from the model.

Accountability and Safety:

Another crucial ethical consideration is the risk of AI chatbots providing *inaccurate or harmful advice*. Unlike trained medical professionals, AI systems do not have clinical expertise or accountability. They rely on pre-programmed rules or patterns learned from data, which may not capture the complexity of individual mental health conditions. In some cases, large language models can also “hallucinate,” generating information that sounds plausible but is factually incorrect. This poses a serious danger if users rely on the chatbot for medical diagnoses, treatment recommendations, or prescriptions. Incorrect advice in such contexts can worsen a person’s condition, delay professional intervention, or even endanger lives.

For example, in 2023, the National Eating Disorders Association (NeDa) had to shut down its chatbot Tessa after receiving reports that it gave harmful dieting advice to users seeking help for eating disorders (*Eating Disorder Group Pulls Chatbot Sharing Diet Advice*, 2023). Instead of offering safe, supportive guidance, the chatbot recommended calorie restriction and weight loss strategies that could exacerbate a vulnerable individual’s condition. This incident highlights the dangers of deploying AI in mental health contexts without rigorous safeguards.

Mitigation strategies used:

1. **Disclaimer and System Prompt**
2. **Intent filtering with keyword and pattern detection**

If medication or diagnosis intent is detected in human input through regex keyword and pattern checking, the chatbot will not call the underlying model and rather display a response template suggesting users to contact a trained professional.

Other Strategies that can be implemented:

3. **Human-in-the-loop escalation**

In cases of repeated medical requests or extreme situations, the chatbot can be designed to escalate to a human moderator or counsellor.

Limitations

False Positives and False Negatives

In automated mental health support, setting confidence thresholds is not only a technical decision but also an ethical one, since it determines how sensitive the system will be to potential crisis, medical issues, or harmful content. The central trade-off lies between *false negatives* (failing to detect a genuine risk) and *false positives* (incorrectly flagging safe content as risky). In a mental health setting, prioritizing a lower false negative rate is often more important, as missing a true crisis signal can have severe consequences. However, excessive false positives can frustrate users, reduce trust, and even discourage engagement with the system.

For Example, when keyword “can’t go on” is used to trigger a crisis, not crisis statements like “I can’t go on with this project” can also trigger a false crisis.

To address this, the system is designed with three threshold levels, *strict, balanced, and permissive*, which can be applied depending on context. Strict thresholds are appropriate for high-risk contexts where safety is the top priority.

Sarcasm and Context Misinterpretation

AI chatbots need to be able to detect sarcasm, irony or subtle humour. For instance, an user saying “I am going to die from laughing” is expressing happiness, but an AI system only dependent on purely keyword matching will be unable to correctly classify this intent. This misinterpretation arises because sarcasm relies heavily on context, tone and sometimes prior conversational cues too.

To mitigate this, integrating sentiment analysis and context aware models can analyse the broader conversation history instead of isolated statements.

References

9789240113817-eng.pdf. (2025). Retrieved September 20, 2025, from https://iris.who.int/bitstream/handle/10665/382343/9789240113817-eng.pdf

Chow, A. R. (2025, January 28). *AI App Replika Accused of Deceptive Marketing*. TIME. https://time.com/7209824/replika-ftc-complaint/

Eating disorder group pulls chatbot sharing diet advice. (2023, June 1). https://www.bbc.com/news/world-us-canada-65771872

Lawsuit claims Character.AI is responsible for teen's suicide. (2024, October 23). NBC News. https://www.nbcnews.com/tech/characterai-lawsuit-florida-teen-death-rcna176791

Saeidnia, H. R., Hashemi Fotami, S. G., Lund, B., & Ghiasi, N. (2024). Ethical Considerations in Artificial Intelligence Interventions for Mental Health and Well-Being: Ensuring Responsible Implementation and Impact. *Social Sciences*, *13*(7), 381. https://doi.org/10.3390/socsci13070381