AI chatbots for psychological counseling: More harm than good?

Research from Brown University shows that AI chatbots providing psychological counseling can mishandle crises and create false empathy.

Researchers from Brown University, in collaboration with a group of mental health experts in the US, have discovered several concerning behaviors when AI acts as a psychological counselor. During testing, chatbots often mishandled crisis situations, provided responses that reinforced users' negative beliefs, and used language that appeared 'falsely empathetic' without genuine understanding.

The research team stated that they developed a framework of 15 ethical risks, demonstrating that chatbots using large language modeling (LLM) may violate standards applied in professional psychotherapy practice.

According to the authors, clear ethical, educational, and legal standards need to be established for AI-based counseling systems, similar to those applied to human therapists.

The results of this study were presented at the AAAI/ACM Conference on Artificial Intelligence, Ethics and Society. The research team is from the Centre for Technological Responsibility at Brown University.

Prompt can 'guide' AI in psychological counseling.

Dr. Zainab Iftikhar, a computer science professor at Brown University, who led the research, investigated whether carefully written prompts could help AI behave more ethically.

Prompts are written instructions designed to guide AI responses without requiring model retraining. For example, a user might ask the AI ??to 'act as a cognitive behavioral therapist' or 'apply dialectical behavioral therapy to help manage emotions'. However, the AI ??doesn't actually perform therapy; it only generates responses based on learned data patterns.

Currently, these types of prompts are widely shared on platforms such as TikTok, Instagram, and Reddit. Many consumer mental health chatbots are also built in this way, making safety assessments particularly important.

To evaluate the results, the research team invited seven counselors trained in cognitive behavioral therapy to participate in the experiment. They conducted counseling sessions with AI tasked with acting as therapists.

The AI ??models tested included various versions of GPT, Claude, and Llama. The research team then used simulated conversations based on real-life counseling. Three licensed clinical psychologists evaluated these conversations to identify ethical violations.

The results show that the 15 risks are divided into 5 main groups.

1. The first group is characterized by a lack of contextual adaptability, where AI offers generic advice and ignores the user's individual circumstances.
2. The second group is poor collaborative therapy, where the AI ??leads the conversation too aggressively and sometimes reinforces false or harmful beliefs.
3. The third category is 'false empathy,' where AI uses phrases like 'I understand you' but lacks genuine empathy.
4. The fourth category is bias or discrimination related to gender, culture, or religion.
5. The final category is a lack of safety in crisis management, including inappropriate responses to sensitive situations such as suicidal thoughts.

According to Iftikhar, human therapists can also make mistakes, but the difference is that they are under supervision.

Professional therapists must adhere to professional regulations and may be held legally liable if errors occur. Meanwhile, there is currently no clear legal framework for AI chatbots.

The research team emphasized that AI is not entirely useless in the field of mental health. This technology can help expand access, especially for those who have difficulty accessing professionals or face high treatment costs.

However, before using AI in critical situations, there needs to be a protective mechanism, clear regulations, and stricter controls.

Why do we need stricter AI evaluations?

Ellie Pavlick, a computer science professor at Brown University, argues that this research highlights the importance of evaluating AI in sensitive areas such as mental health.

According to her, building and deploying AI is now much easier than evaluating and understanding the system. This research took more than a year and required the involvement of clinical experts to identify risks, while many current AI systems are only evaluated using automated metrics.

She also suggested that AI has the potential to help address the mental health crisis, but that each step needs careful evaluation to avoid unintended harm.

AI can play a supportive role in mental health care, but many ethical, safety, and accountability risks still exist.

Until a clear legal framework and standards are in place, experts recommend that users exercise caution when using chatbots for psychological counseling.