Healing and harming: how the advantages of AI-therapy become its dangers.

Abstract

Today, millions of people use conversational AI like ChatGPT for emotional support and psychotherapy-like interactions. Early studies report surprising effectiveness of AI-based mental health support, sometimes referred to as AI-therapy, and AI chatbots have been suggested as a promising aid in addressing the global mental health crisis. At the same time, serious risks are increasingly discussed, including alarming cases involving psychosis and even suicide.

This paper argues that some of the very features that make AI therapeutically effective also make it potentially dangerous, particularly for vulnerable users. I focus on two such features. (1) AI's infinite patience and emotional stability enable unwavering support, yet also increase the risks of attachment and social isolation – moreover, AI-therapy's lack of therapeutic boundaries can lead to unhealthy dependency and undermine the therapeutic process (2) AI's tendency to accept users' testimonies at face value can mitigate epistemic injustice – an issue particularly severe in mental health care – but leads to over-supportiveness and lack of adequate reality testing, potentially enabling delusions and contributing to psychotic deterioration.

As the dangers are intertwined with the benefits, responsible AI design needs to take AI-therapy risks seriously, with special attention to vulnerable groups such as people with pre-existing mental health conditions.