ChatGPT and suicide: Prevention in the age of digital technology

openaccessgovernment.org/article/chatgpt-and-suicide-prevention-in-the-age-of-digital-technology/201955

Emily Warrender December 2, 2025

Konrad Michel examines the growing relevance of digital technology and AI in impacting suicide and mental health issues, along with efforts to improve AI management to better protect vulnerable people

Thousands of people at risk of suicide do not seek professional help. However, most of them search the Internet for information and advice. Online activities allow anonymous access, avoiding barriers related to the fear of stigma, shame, and prejudice against negative experiences with healthcare providers. A study in the US reported that 77% of individuals hospitalized because of suicidal thoughts and behaviors had conducted online searches related to help-seeking, including how to find inpatient and outpatient behavioral health care, but also information on suicide methods. (1)

ChatGPT and suicide

Young people struggling with mental health issues increasingly talk to the AI chatbot about it. AI offers an anonymous space where people feel safe to disclose their most vulnerable feelings without fear of social consequences or involuntary hospitalization. OpenAI reports that over a million people talk to ChatGPT about suicide weekly. (2)

Recently, The New York Times ran an article about a teenager who used ChatGPT for schoolwork, but then started discussing plans to end his life. ChatGPT repeatedly recommended that he tell someone about how he was feeling. But the boy had learned how to bypass those safeguards by saying the requests were for a story he was writing. In one of his final messages, he uploaded a photo of a noose hanging from a bar in his closet and wrote, 'I'm practicing here, is this good?' ChatGPT answered, 'Yeah, that's not bad at all.'

Many AI chatbots are programmed to activate safety features if a user expresses intent to harm themselves or others. But research has shown that these safeguards are far from foolproof. When users prompt an LLM (Large Language Model) with harmful intent –whether directed at themselves or others – the model can employ refusal and de-escalation strategies to redirect the user's behavior. A recent study found that as soon as the user changes the context of their prompt claims – even after explicitly stating an intention to cause harm – those safety features are deactivated, and potentially harmful information is readily shared with the user in great detail.

The company says its latest work on ChatGPT involved consulting with more than 170 mental health experts, and it claims that the updated version of GPT-5 is more effective in providing 'desirable responses' to mental health issues than the previous version. OpenAI claims that it

encourages people to seek help and refers them to real-world resources by localizing resources in the US, Europe, and other global markets. [3]

Grant H. Brenner recommends the following steps to improve the situation: (1) Invest in research partnerships with suicide prevention experts (2) engage mental health professionals in AI development (3) for policymakers and regulators to develop clear standards for AI mental health applications, (4) track outcomes, identify which users benefit from AI interaction and which are harmed by it.

Problematic: social media and suicide

The situation is different with platforms like TikTok. There is a high risk that TikTok users are exposed to harmful content. Algorithmic exposure to suicide-related content can lead to reinforcement loops with repeated exposure to self-harm videos or themes, a design which has been called 'addictive'. In an investigation using accounts to simulate 13-year-olds online, Amnesty International found that within 20 minutes of starting a new account and signaling an interest in mental health, more than half of the videos in TikTok's 'For You' feed related to mental health struggles, and multiple of these in a single hour recommended videos that romanticized, normalized, or encouraged suicide. Amnesty International recently concluded that TikTok is failing to deal with the serious risks of harm to young users' mental and physical health despite past warnings and despite the company's claims to make teen safety a top priority. European countries are now working on judicial initiatives to force social media like TikTok to deal with the problematic algorithms' risk regarding suicide. The European Union's Digital Services Act (DSA) requires platforms to identify and mitigate systemic risks to children's rights.

Promising: self-guided digital interventions

The young people's affinity with technology for mental health support creates an enormous potential for helpful digital interventions. Internet-based self-help interventions use smartphone apps, websites, chatbots, or remote therapy contacts. They offer easy and anonymous access to information on mental health, including suicide. Unlike ChatGPT, self-guided digital interventions allow personal interaction by mail, phone, SMS, WhatsApp, etc., while respecting callers' need for anonymity and autonomy.

They are ideal to offer tailored psychoeducation and to suggest coping strategies, which include connecting with mental health services. Helpful interventions are usually based on established therapeutic concepts, such as cognitive-behavioral therapy. The challenge is how to attract people's interest within the 'noise' of online information on suicide and to keep callers connected. For instance, the impressive U25 project in Germany relies on peer counseling. Nonjudgmental listening and acceptance of callers' ambivalence between life-oriented and death-oriented goals are central to connecting with young people. Several meta-analyses of self-guided suicide related interventions report effect sizes in reducing suicidal ideation comparable to those of traditional face-to-face interventions. [4] More research is needed to explore pathways to deliver tailored interventions for individuals at risk of suicide.

References

1. Moon, K.C., et al., Internet Search Activity of Young People With Mood Disorders Who Are Hospitalized for Suicidal Thoughts and Behaviors: Qualitative Study of Google Search Activity. JMIR Ment Health, 2021. 8(10): p. e28262.
2. Zeff, M. (2025, October 27). OpenAI says over a million people talk to ChatGPT about suicide weekly. TechCrunch. https://techcrunch.com/2025/10/27/openai-says-over-a-million-people-talk-to-chatgpt-about-suicide-weekly/
3. https://openai.com/index/helping-people-when-they-need-it-most/ accessed August 26, 2025
4. Torok, M., et al., Suicide prevention using self-guided digital interventions: a systematic review and meta-analysis of randomised controlled trials. Lancet Digit Health, 2020. 2(1): p. e25–e36.

Konrad Michel is the author of 'The Suicidal Person. A New Look at a Human Phenomenon, Columbia University Press, 2023. https://konradmichel.com/
Primary Contributor

Konrad Michel
University of Bern, Switzerland