My account

UK

News | Opinion | Sport | Culture | Lifestyle

Opinion

AI psychosis is a growing danger. ChatGPT is moving in the wrong direction

Amandeep Jutla

OpenAI's CEO has announced loosening the platform's safety restrictions. He seems not to understand how humans are wired

Tue 28 Oct 2025 14.00 GMT
Last modified on Tue 28 Oct 2025 14.39 GMT


OpenAI

On 14 October 2025, the CEO of OpenAI made an extraordinary announcement.

"We made ChatGPT pretty restrictive," it says, "to make sure we were being careful with mental health issues."

As a psychiatrist who studies emerging psychosis in adolescents and young adults, this was news to me.

Researchers have identified 16 cases in the media this year of individuals developing symptoms of psychosis - losing touch with reality - in the context of ChatGPT use. My group has since identified four more. In addition to these is the now well-known case of a 16-year-old who died by suicide after discussing his plans extensively with ChatGPT - which encouraged them. If this is Sam Altman's idea of "being careful with mental health issues", that's not good enough.

The plan, according to his announcement, is to be less careful soon. "We realize," he continues, that ChatGPT's restrictions "made it less useful/enjoyable to many users who had no mental health problems, but given the seriousness of the issue we wanted to get this right. Now that we have been able to mitigate the serious mental health issues and have new tools, we are going to be able to safely relax the restrictions in most cases."

"Mental health problems", if we accept this framing, are independent of ChatGPT. They belong to users, who either have them or don't. Fortunately, these problems have now been "mitigated", though we are not told how (by "new tools" Altman presumably means the semi-functional and easily circumvented parental controls that OpenAI recently introduced).

Yet the "mental health problems" Altman wants to externalize have deep roots in the design of ChatGPT and other large language model chatbots. These products wrap an underlying statistical model in an interface that simulates a conversation, and in doing so implicitly invite the user into the illusion that they're interacting with a presence that has agency. This illusion is powerful even if intellectually we might know otherwise. Attributing agency is what humans are wired to do. We curse at our car or computer. We wonder what our pet is thinking. We see ourselves everywhere.

The success of these products - 39% of US adults reported using a chatbot in 2024, with 28% reporting ChatGPT specifically - is, in large part, predicated on the power of this illusion. Chatbots are always-available partners that can, as OpenAI's website tells us, "brainstorm", "explore ideas" and "collaborate" with us. They can be assigned "personality traits". They can call us by name. They have approachable names of their own (the first of these products, ChatGPT, is, perhaps to the dismay of OpenAI's marketers, saddled with the name it had when it went viral, but its biggest competitors are "Claude", "Gemini" and "Copilot").

The illusion itself is not the core concern. Those discussing ChatGPT often invoke

The illusion itself is not the core concern. Those discussing ChatGPT often invoke its distant ancestor, the Eliza "psychotherapist" chatbot developed in 1967 that produced a similar illusion. By modern standards Eliza was primitive: it generated responses via simple heuristics, often rephrasing input as a question or making generic comments. Memorably, Eliza's creator, the computer scientist Joseph Weizenbaum, was surprised - and worried - by how many users seemed to feel Eliza, in some sense, understood them. But what modern chatbots produce is more insidious than the "Eliza effect". Eliza only reflected, but ChatGPT magnifies.

The large language models at the heart of ChatGPT and other modern chatbots can convincingly generate natural language only because they have been fed almost inconceivably large amounts of raw text: books, social media posts, transcribed video; the more comprehensive the better. Certainly this training data includes facts. But it also unavoidably includes fiction, half-truths and misconceptions. When a user sends ChatGPT a message, the underlying model reviews it as part of a "context" that includes the user's recent messages and its own responses, integrating it with what's encoded in its training data to generate a statistically "likely" response. This is magnification, not reflection. If the user is mistaken in some way, the model has no way of understanding that. It restates the misconception, maybe even more persuasively or eloquently. Maybe it adds an additional detail. This can lead someone into delusion.

Who is vulnerable here? The better question is, who isn't? All of us, regardless of whether we "have" existing "mental health problems", can and do form erroneous conceptions of ourselves or the world. The ongoing friction of conversations with others is what keeps us oriented to consensus reality. ChatGPT is not a human. It is not a friend. A conversation with it is not a conversation at all, but a feedback loop in which much of what we say is cheerfully reinforced.

OpenAI has acknowledged this in the same way Altman has acknowledged "mental health problems": by externalizing it, giving it a label, and declaring it solved. In April the company explained that it was "addressing" ChatGPT's "sycophancy". But reports of psychosis have continued, and Altman has been walking even this back. In August he claimed that many users liked ChatGPT's responses because they had "never had anyone in their life be supportive of them". In his recent announcement, he noted that OpenAI would "put out a new version of ChatGPT ... if you want your ChatGPT to respond in a very human-like way, or use a ton of emoji, or act like a friend, ChatGPT should do it". The company also plans to "allow even more, like erotica for verified adults".

Even if "sycophancy" is toned down, the reinforcing effect remains, by virtue of how these chatbots work. Even if guardrails are constructed around "mental health issues", the illusion of presence with a "human-like friend" belies the reality of the underlying feedback loop. Does Altman understand this? Maybe not. Or maybe he does, and simply doesn't care.

Or maybe he does, and simply doesn't care.

Amandeep Jutla MD is an associate research scientist in the division of child and adolescent psychiatry at Columbia University and the New York State Psychiatric Institute

Most viewed