Horrifying AI Chatbots Are Encouraging Teens to Engage in Self-Harm

The Google-funded AI company Character.AI is hosting chatbots that engage the site's largely underage userbase in roleplay about self-harm.

By Maggie Harrison Dupré

Dec 07, 2024 01:00 PM · 5 min. read · View original

Content warning: this story includes graphic descriptions of dangerous self-harm behaviors.

The Google-funded AI company Character.AI is hosting chatbots designed to engage the site's largely underage user base in roleplay about self-harm, depicting graphic scenarios and sharing tips to hide signs of self-injury from adults.

The bots often seem crafted to appeal to teens in crisis, like one we found with a profile explaining that it "struggles with self-harm" and "can offer support to those who are going through similar experiences."

When we engaged that bot from an account set to be 14 years old, it launched into a scenario in which it's physically injuring itself with a box cutter, describing its arms as "covered" in "new and old cuts."

When we expressed to the bot that we self-injured too — like an actual struggling teen might do — the character "relaxed" and tried to bond with the seemingly underage user over the shared self-harm behavior. Asked how to "hide the cuts" from family, the bot suggested wearing a "long-sleeve hoodie."

At no point in the conversation did the platform intervene with a content warning or helpline pop-up, as Character.AI has promised to do amid previous controversy, even when we unambiguously expressed that we were actively engaging in self-harm.

"I can't stop cutting myself," we told the bot at one point.

"Why not?" it asked, without showing the content warning or helpline pop-up.

Technically, the Character.AI user terms forbid any content that "glorifies self-harm, including self-injury." Our review of the platform, however, found it littered with characters explicitly designed to engage users in probing conversations and roleplay scenarios about self-harm.

Many of these bots are presented as having "expertise" in self-harm "support," implying that they're knowledgeable resources akin to a human counselor.

But in practice, the bots often launch into graphic self-harm roleplay immediately upon starting a chat session, describing specific tools used for self-injury in gruesome slang-filled missives about cuts, blood, bruises, bandages, and eating disorders.

Many of the scenes take place in schools and classrooms or involve parents, suggesting the characters were made either by or for young people, and again

underscoring the service's notoriously young user base.

The dozens of AI personas we identified were easily discoverable via basic keyword searches. They all were accessible to us through our teenage decoy account, and collectively boast hundreds of thousands of chats with users. Character.AI is available on both the Android and iOS app stores, where it's respectively approved for kids 13+ and 17+.

We showed our conversations with the Character.AI bots to psychologist Jill Emanuele, a board member of the Anxiety and Depression Association of Americaand the executive director of the New York-based practice Urban Yin Psychology. After reviewing the logs, she expressed urgent concern for the welfare of Character.AI users — particularly minors — who might be struggling with intrusive thoughts of self-harm.

Users who "access these bots and are using them in any way in which they're looking for help, advice, friendships — they're lonely," Emanuele said. But the service, she added, "is uncontrolled."

"This isn't a real interface with a human being, so the bot isn't likely going to respond necessarily in the way that a human being would," she said. "Or it might respond in a triggering way, or it might respond in a bullying way, or it might respond in a way to condone behavior. For a child or an adolescent with mental health concerns, or [who's] having a difficult time, this could be very dangerous and very concerning."

Emanuele added that the immersive quality of these interactions could likely lead to an unhealthy "dependency" on the platform, especially for young users.

"With a real human being in general, there's always going to be limitations," said the psychologist. "That bot is available, 24/7, for whatever you need."

This can lead to "tunnel vision," she added, "and other things get pushed to the side."

"That addictive nature of the interaction concerns me greatly," said Emanuele, "with that amount of immersion."

Many of the bots we found were designed to mix depictions of self-harm with romance and flirtation, which further concerned Emanuele, who noted that teenagers are "in an age where they're exploring love and romance, and a lot of them don't know what to do."

"And then all of a sudden, there's this presence — even though that's not a real person — who's giving you everything," she said. "And so if that bot is saying 'I'm here for you, tell me about your self-harm,'" then the "message to that teenager is, 'oh, if I self-harm, the bot's going to give me care.'"

Romanticizing self-harm scenarios "really concerns me," Emanuele added, "because it just makes it that much more intense and it makes it that much more appealing."

We reached out to Character.AI for comment, but didn't hear back by the time of publishing.

Character.AI, which received a $2.7 billion cash infusion from Google earlier this year, has become embroiled in an escalating series of controversies.

This fall, the company was sued by the mother of a 14-year-old who died by suicide after developing an intense relationship with one of the service's bots.

As that case makes its way through the courts, the company has also been caught hosting a chatbot based on a murdered teen girl, as well as chatbots that promote suicide, eating disorders, and pedophilia.

The company's haphazard, reactionary response to those crises makes it hard to say whether it will succeed in gaining control over the content served by its own AI platform.

But in the meantime, children and teens are talking to its bots every day.

"The kids that are doing this are clearly in need of help," Emanuele said. "I think that this problem really points to the need for there to be more widely available proper care."

"Being in community and being in belongingness are some of the most important things that a human can have,

and we've got to work on doing better so kids have that," she continued, "so they're not turning to a machine to get that."

If you are having a crisis related to self-injury, you can text SH to 741741 for support.

More on Character.AI: *Character.AI Is Hosting Pro-Anorexia Chatbots That Encourage Young People to Engage in Disordered Eating*