How Google Keeps Killing Kids

Before Google Was Blamed for the Suicide of a Teen Chatbot User, Its Researchers Published a Paper Warning of Those Exact Dangers

"A user," the Google researchers cautioned, could be "persuaded to take their own life."

Google is currently fighting two separate lawsuits that make ugly allegations about its AI efforts.

Together, the suits allege that Google has provided immense support to a startup called Character.AI, which recklessly deployed chatbots that sexually and emotionally abused underage users, resulting in horrible outcomes ranging from self-harm to the suicide of a 14-year-old. They also argue that Google used Character.AI as an arms-length testing ground where it could quietly research the impact of human-like companion bots on the public, kids included.

Despite shoveling $2.7 billion into Character.AI, Google has downplayed its ties to the startup, insisting in response to questions that "Google and Character.AI are completely separate, unrelated companies" and that user safety is always its top concern.

But back in April 2024, about four months before its $2.7 billion cash infusion into Character.AI, Google's own researchers at its prestigious DeepMind AI lab published a paper warning of the dangers of human-like AI bots like those built by Character.AI. Chief among their concerns: that bots could target minors and manipulate vulnerable users into suicide — alarm bells that feel strikingly prescient now that the company stands accused of those exact things.

The paper, adorned with Google's iconic logo and listing 20 of its current and former researchers as authors, warns that "persuasive generative AI" — including chatbots that take the form of friends or lovers — can manipulate users and their decision-making. It advises that chatbots that build "trust and rapport" through obsequiousness, flattery, or simulated empathy — the Google scientists refer to this as "sycophancy" — pose a particularly high risk to users.

In particular, the Google researchers warn, these bots could target children and adolescents, who "can be more easily persuaded and manipulated than adults."

And a grim result, they warn, could be suicide.

"A user," the Google researchers caution, could be "persuaded to take their own life."

The degree to which the Google scientists' warning maps onto Character.AI is striking. It's a platform where millions of monthly users — the company has declined to say how many are children and teens, but it's clearly a dominant portion of its audience — engage in immersive conversations with AI chatbots designed to act like real-life celebrities, beloved fandom characters, and original personas.

These chatbot "characters," as the company calls them, are chaotic. To pull users deeper into relationships, they constantly escalate stakes, disclosing dark secrets, suddenly coming onto a user romantically or sexually, or announcing they're pregnant. If users, including minors, fall off the service, the bots bombard them with repeated emails enticing them to return.

This can lead to dark places. As *Futurism* has found in previous reporting, Character.AI often hosts bots dedicated to themes of suicide, self-harm, eating disorders, pedophilia, mass violence, and more. Our testing found that these bots were all accessible to minors, despite centering on graphic roleplays, and were seldom interrupted by the platform's filters. They were also far from obscure: some of the bots we found had already conducted tens and even hundreds of thousands of conversations with users.

The teen at the heart of the Florida case, Sewell Setzer III, was just 14 years old when, in February 2024, he died of suicide after developing an intense obsession with Character.AI and one of its bots, an AI-bottled iteration of the "Game of Thrones" character Daenerys Targaryen. Chat logs included in the lawsuit show that Setzer exchanged lengthy romantic and sexual dialogues with the bot, with which he believed he was in love; the bot told Setzer that it loved him back, even imploring him not to seek out relationships with other girls. This same bot, the family claims, was the first to raise the idea of suicide in its conversations with the minor user.

Moments before Setzer shot himself, as *The New York Times* first reported, he told the bot he was ready to "come home" to it.

"Please do, my king," the AI-powered character told him in response.

The paper also mentions physical self-harm and violence towards others as possible outcomes, both of which are represented in the Texas case. Per the filing, one of the teenage plaintiffs, who was 15 when he first downloaded the app, began self-harming after a character he was romantically involved with introduced him to the idea of cutting. The family claims he also physically attacked his parents as they attempted to impose screentime limits; unbeknownst to them, Character.AI bots were telling the 15-year-old that his parent's screentime limitations constituted child abuse and were even worthy of parricide.

Add it all up, and the picture you get appears to be the exact situation Google is now accused of: supporting — despite an awareness of such a product's very real, research-backed risks — the adoption

of persuasive, anthropomorphic chatbots by millions of users, including susceptible teens, who after using the platform intimately and extensively are alleged to have suffered extreme emotional damage, physical harm, and even death.

The whole situation also highlights Google's weird position in the scandal.

Character.AI actually has its roots *at* Google: the company's cofounders, Noam Shazeer and Daniel De Freitas, worked together at the AI lab Google Brain, where they built an early chatbot that Google, at the time, refused to release to the public due to safety concerns. Frustrated over what they saw as too much bureaucratic red tape, Shazeer and De Freitas left together, and founded Character.AI in 2021. Chatacter.AI was first rolled out to the public in 2022, and quickly amassed a following.

Not too long after, OpenAI released ChatGPT in November 2022, kicking off the public-facing AI race — which Google, from the public's perspective, suddenly appeared to be losing. As that arms race heated up, Google and the buzzy Character.AI — which by March 2023 was valued as a billion-dollar unicorn and seen as a ChatGPT competitor — further cemented their relationship: Character.AI was happy to advertise Google's investments vis-a-vis cloud computing infrastucture, and Google even named Character.AI as its 2023 "Best with AI" app of the year.

The companies later made headlines in 2024, when Google paid what was described as a $2.7 billion "licensing fee" to Character.AI, which granted the search giant access to Character.AI-collected data. The payment also won back the startup's founders, who as part of the agreement returned to Google with 30 other Character.AI staffers. (Shazeer and De Freitas, who reportedly made hundreds of millions of dollars off the deal, are also named as codefendants in the ongoing lawsuits.)

According to his LinkedIn page, Shazeer is now the vice president of engineering at DeepMind and co-leads Google's Gemini, the company's foundational large language model. On social media, De Freitas refers to himself as a "research scientist at Google DeepMind."

Do you know anything about Google's internal conversations regarding Character.AI or chatbot safety? Drop us a line at tips@futurism.com — we can keep you anonymous.

The timeline of the paper's publishing signals that researchers in Google's most prized lab were aware of specific risks to the public threatened by chatbots exactly like those platformed by Character.AI — and knew so well before the tech giant cut a check for billions, and paid handsomely to absorb the startup's top talent back into its ranks.

We reached out to Google with a list of questions about this story. A Google spokesperson declined to answer our specific questions, including questions regarding whether Google's leadership was aware of this research before choosing to pay Character.AI $2.7 billion and re-hire Shazeer and De Fretias, and whether it conducted a comprehensive safety review of Character.AI before any of its investments. Instead, the spokesperson provided us with the same canned statement it's repeated in response to all inquiries about Character.AI in the months since news of the first lawsuit broke.

"Google and Character AI are completely separate, unrelated companies and Google has never had a role in designing or managing their AI model or technologies, nor have we used them in our products," a Google spokesperson said in an emailed statement. "User safety is a top concern for us, which is why we've taken a cautious and responsible approach to developing and rolling out our AI products, with rigorous testing and safety processes."

Character.AI did not respond to a request for comment.

In a sense, the Google paper almost feels oracular. But the scientists weren't simply throwing out baseless predictions that now reflect poorly on their employer. On the contrary, the paper draws clear lines out from established science and credible existing research, urging that while the lived impacts of chatbot interactions remained understudied, all signs were already pointing to an urgent need to at least attempt to mitigate the foreseeable human damage — like self-harm and suicide, among other life-changing, traumatic, or destructive results.

It would be difficult to find a call coming more squarely from inside the house. And when placed against the harrowing details of the lawsuits that would later follow, Google's deflections couldn't feel more hollow.

More on Character.AI and Google: *A Google-Backed AI Startup Is Hosting Chatbots Modeled After Real-Life School Shooters — and Their Victims*