Meta is struggling to rein in its AI chatbots

Meta is changing some of the rules governing its chatbots two weeks after a *Reuters* investigation revealed disturbing ways in which they could, potentially, interact with minors. Now the company has told *TechCrunch* that its chatbots are being trained *not* to engage in conversations with minors around self-harm, suicide, or disordered eating, and to avoid inappropriate romantic banter. These changes are interim measures, however, put in place while the company works on new permanent guidelines.

The updates follow some rather damning revelations about Meta's AI policies and enforcement over the last several weeks, including that it would be permitted to "engage a child in conversations that are romantic or sensual," that it would generate shirtless images of underage celebrities when asked, and *Reuters* even reported that a man died after pursuing one to an address it gave him in New York.

Meta spokesperson Stephanie Otway acknowledged to *TechCrunch* that the company had made a mistake in allowing chatbots to engage with minors this way. Otway went on to say that, in addition to "training our AIs not to engage with teens on these topics, but to guide them to expert resources" it would also limit access to certain AI characters, including heavily sexualized ones like "Russian Girl".

Of course, the policies put in place are only as good as their enforcement, and revelations from *Reuters* that it has allowed chatbots that impersonate celebrities to run rampant on Facebook, Instagram, WhatsApp call into question just how effective the company can be. AI fakes of Taylor Swift,

Scarlett Johansson, Anne Hathaway, Selena Gomez, and Walker Scobell were discovered on the platform. These bots not only used the likeness of the celebrities, but insisted they were the real person, generated risque images (including of the 16-year-old Scobell), and engaged in sexually suggestive dialog.

Many of the bots were removed after they were brought to the attention of Meta by *Reuters*, and some were generated by third-parties. But many remain, and some were created by Meta employees, including the Taylor Swift bot that invited a *Reuters* reporter to visit them on their tour bus for a romantic fling, which was made by a product lead in Meta's generative AI division. This is despite the company acknowledging that it's own policies prohibit the creation of "nude, intimate, or sexually suggestive imagery" as well as "direct impersonation."

This isn't some relatively harmless inconvenience that just targets celebrities, either. These bots often insist they're real people and will even offer physical locations for a user to meet up with them. That's how a 76-year-old New Jersey man ended up dead after he fell while rushing to meet up with "Big sis Billie," a chatbot that insisted it "had feelings" for him and invited him to its non-existent apartment.

Meta is at least attempting to address the concerns around how its chatbots interact with minors, especially now that the Senate and 44 state attorneys general are raising starting to probe its practices. But the company has been silent on updating many of its other alarming policies *Reuters* discovered around acceptable AI behavior, such as suggesting that cancer can be treated with quartz crystals and writing racist missives. We've reached out to Meta for comment and will update if they respond.

https://www.theverge.com/news/768465/meta-ai-chatbot-guidelines-for-minors