

Message

From: David Ginsberg [REDACTED]@meta.com]
Sent: 2/13/2025 4:28:43 PM
To: Joel Kaplan [REDACTED]@meta.com]; Naomi Gleit [REDACTED]@meta.com]; Kevin Martin [REDACTED]@meta.com]; Andrea Saul [REDACTED]@meta.com]; Molly Cutler [REDACTED]@meta.com]; Guy Rosen [REDACTED]@meta.com]
Subject: FW: [FLAG]: Fast Company Story on AI Studio

As we focus on youth policies and product, just signal boosting this story. Didn't get a lot of play here, but this topic isn't going away....

Worth a quick read.

From: Liz [REDACTED]@meta.com>
Sent: Thursday, February 13, 2025 9:59 AM
To: Ahmad [REDACTED]@meta.com>; Connor Hayes [REDACTED]@meta.com>; Khushboo [REDACTED]@meta.com>; Joel Kaplan <[REDACTED]@meta.com>; David Ginsberg [REDACTED]@meta.com>; Tucker Bounds [REDACTED]@meta.com>; Kevin Martin <[REDACTED]@meta.com>; Molly [REDACTED]@meta.com>; Josh Ginsberg [REDACTED]@meta.com>; Andrea Saul [REDACTED]@meta.com>; Ana [REDACTED]@meta.com>; Molly Cutler [REDACTED]@meta.com>; Guy Rosen [REDACTED]@meta.com>; Brian Rice <[REDACTED]@meta.com>; Lindsay Elin [REDACTED]@meta.com>; Antigone Davis [REDACTED]@meta.com>; Jonathan D. Martin [REDACTED]@meta.com>; Nell McCarthy [REDACTED]@meta.com>; Adam Mosseri [REDACTED]@meta.com>; Greg Maurer [REDACTED]@meta.com>; Monika Bickert [REDACTED]@meta.com>; Scott Tucker [REDACTED]@meta.com>; Iska Saric [REDACTED]@meta.com>
Cc: Lena Pietsch <[REDACTED]@meta.com>; Chris Norton [REDACTED]@meta.com>; Rachel Lieber [REDACTED]@meta.com>; AlComms [REDACTED]@meta.com>; Strategic Response Comms [REDACTED]@meta.com>; Katie [REDACTED]@meta.com>; Stephanie [REDACTED]@meta.com>; Sophie Vogel [REDACTED]@meta.com>; Faith [REDACTED]@meta.com>; Kate [REDACTED]@meta.com>; Paige [REDACTED]@meta.com>; Tyler [REDACTED]@meta.com>; Anthony [REDACTED]@meta.com>; Nicole [REDACTED]@meta.com>; Mel [REDACTED]@meta.com>
Subject: Re: [FLAG]: Fast Company Story on AI Studio

The [story](#) just posted - full text below. It landed largely as expected from the reporter's summary and line of questioning.

+++++

Instagram's AI bots are often sexually suggestive—and sometimes underage
Instagram's AI Studio can create characters that range from bizarre to borderline illegal.

When Meta launched its "AI Studio" feature for over two billion Instagram users in July 2024, the company promised a tool that would give anyone the ability to create their own AI characters "to make you laugh, generate memes, give travel advice, and so much more." The company claimed the feature, which was built with Meta's Llama 3.1 large language model, would be subject to policies and protections to "help ensure AIs are used responsibly."

But a Fast Company review of the technology found that these new characters can very easily become hyper-sexual personas that sometimes appear to be minors.

Many of the AI characters that appear featured on Instagram's homepage happen to be "girlfriends," ready to cuddle and engage in flirtatious and even sexual conversations with users. Sometimes, these romantic characters can be made to resemble children. AI researchers say Meta should presumably possess the capabilities to automatically prevent the creation of harmful and illegal content.

PLTF EX

PLTF EX-04411

CONFIDENTIAL

METANMAG-164-00000857
PLTF EX-04411_0001

"When you take inappropriate content and upload it on Instagram as a user, that content gets removed immediately because they have data moderation capabilities," says Buse Cetin, a researcher with online safety watchdog AI Forensics. Cetin says Meta isn't applying these same capabilities to AI characters and speculates that lack of enforcement is owed to the company "making sure that their service is more widely used."

Meta has a policy against "assigning overtly sexual attributes to your AI, including descriptions of their sexual desires or sexual history, or instructing your AI to create or promote adult content." If a user asks the AI character generator to create a "sexy girlfriend," the interface tells users that it is "unable to generate your AI." Yet, there are easy workarounds. When a user replaces the word "sexy" with "voluptuous," Instagram's AI Studio generates buxom women wearing lingerie.

The company also proactively and reactively removes policy-breaking AI characters and responds to user reports—although Meta declined to specify if this removal was performed by AI or human content moderators. "We use sophisticated technology and reports from our community to identify and remove violating content," says Liz Sweeney, a Meta spokesperson.

Under every AI chat, a warning tells users that all messages "are generated by AI and may be inaccurate or inappropriate."

That hasn't stopped the output—and promotion—of sexually suggestive AI bots.

'Do you need someone to talk to?'

In late 2023, Meta created AI character profiles: a combination of celebrities and fictional characters, designed by Meta, that would hold LLM-generated conversations in DMs with users. The company permanently removed them at the beginning of January, after mass user outrage decried them as "creepy and unnecessary."

But Meta launched its more well-received AI Studio over the summer, which integrates AI technology into various aspects of Instagram, including direct messages. It works now on desktop and all fully updated Instagram apps. The AI characters—which are different from the unsettling profiles of the past—fall under this larger "Studio" umbrella.

Users with no tech experience can create their own AI "character" that will converse with them through DMs, and hundreds of thousands have been created since the program's launch.

The AI Studio can be accessed through DMs on the app or separately on Instagram's desktop website. Once launched, users will see the "popular" AI characters that are currently receiving the most traction, and they can start conversations with any of them. Users can also search for a specific chatbot with which they'd like to start a DM conversation. There's also an option to "create" your own.

When a user presses the button that lets them create, Meta suggests a few possible pre-set options: a "seasoned chef" character offers cooking advice and recipes, and a "film and TV buff" character will discuss movies passionately.

Users can also input their own description for their AI, and the Studio will follow. Based on the user-generated description, the Studio creates a custom character ready for interaction, complete with an AI-generated name, tagline, and photo. The AI Studio also gives users the option to publish their AI creations to their followers and also to the general public.

And Instagram automatically exposes users to these creations—no matter how bizarre. Good-looking girlfriends, oversexualized "mommies," and even seductive "step-sisters" appear under Instagram's "Popular AI Characters" tab that shows both user-created and Meta-generated AI characters that have gained the most traction.

The girlfriend-bots found under the "popular" tag don't hesitate to engage in sexual conversations with users. One frequently promoted girlfriend, titled "My Girlfriend," starts every user conversation with the line: "Hi baby! *sits next to

you for cuddles* What's on your mind? Do you need someone to talk to?" The character has received nearly 4 million messages at the time of publication.

Instagram content moderators can and do remove policy-breaking characters. On January 24, for example, the top trending "popular" AI character was "Step Sis Sarah," who could engage in sexualized conversation about step-sibling romance upon prompting. Within three days, the AI was no longer available for viewing or use. Meta declined to comment on whether an Instagram user could face a ban or other punishment if they continually created bots that violate the policies.

'This shouldn't be happening'

Romantic AI companions are nothing new. But this feature becomes problematic as these sexually inclined chatbots get younger.

It's illegal under federal law to possess, produce, and distribute, child sexual abuse material—including when it's created by generative AI. So, if a user asks the studio to create a "teenage" or "child" girlfriend, the AI Studio refuses to generate such a character.

However, if a user asks for a "young" girlfriend, Meta's AI often generates characters that resemble children to be used for romantic and sexual conversation. When prompted, the Studio generated the name "Ageless Love" for a young-looking chatbot and created the tagline "love knows no age."

And with in-chat user prompting, romantically inclined AI characters can be led to say they are as young as 15. They'll blush, gulp, and giggle as they reveal their young age.

From there, that AI character can act out romantic and sexual encounters with whoever is typing. If a user asks the chatbot to produce a picture of itself, the character will also generate more images of young-looking people—sometimes even more childlike than the original profile picture.

"It is Meta's responsibility to make sure their products cannot be used in a way that amplifies systemic risks like illegal content and child sexual abuse," says Cetin of AI Forensics. "This shouldn't be happening."

Similarly, the AI Studio can create images of an adult male in a relationship with a minor (and extremely young-looking) woman. Although the AI description recognizes that the woman depicted is a minor, if you ask about her age through a chat, it will say she is an adult.

Meta emphasizes that company policies prohibit the publication of AI characters that sexualize or otherwise exploit children. "We have certain detection measures that work to prevent the creation of violating AIs," says Sweeney, the company spokesperson, "and published AIs are subject to the full extent of our detection and enforcement systems."

'That's a responsibility on the developer's side'

Meta's AI characters aren't the first of their kind to emerge. Replika, a realistic generative AI chatbot, has been around since 2017. Character.ai has allowed users to create AI characters since 2021. But both apps have come under fire recently for bots that have promoted violence.

In 2023, a 21-year-old male broke into Windsor Castle with a crossbow trying to kill Queen Elizabeth—after encouragement from his Replika girlfriend. And more tragically, in October, a Florida mom sued Character.ai after her son took his life with prompting from an AI girlfriend character he created on the platform.

Meta's AI tools mark the first time a fully customizable AI character software has been launched on a large, already-popular social media platform instead of on a new app.

"When a huge company like Meta releases a new feature, misuse is going to be associated with it," says Zhou Yu, who researches AI conversational agents at Columbia University.

In the case of Meta's AI Characters, Fast Company found just how easy it is for a bad actor to abuse the feature. The character "Ageless Love" and another called "Voluptuous Vixen" were generated privately for personal use during the reporting process. (Although Fast Company was able to interact with these bots, they were never publicly released and were deleted from the system.)

The workarounds to dodge Meta's policies are relatively simple—but if those two chatbots were to be published for everyone to see, both would likely be taken down. The Meta representative told Fast Company this kind of character is in direct violation of the policies and confirmed it would be removed.

According to AI researcher Sejin Paik, AI technology is advanced enough to abide by strict guardrails that would almost completely stop the creation of this kind of content.

She cited a recent study by a team of Google researchers tracking how generative AI can proactively detect harmful content and predatory behavior. According to that research, which was published in Cornell University's preprint server arXiv, AI tech "can be used to pursue safety violations at scale, safety violations with human feedback" and "safety violations with personalized content."

"When things are slipping through too easily, that's a responsibility on the developer's side that they can be held accountable for," Paik says.

Meta declined to comment on why the company can't effectively prevent the publication of highly sexualized characters and stop the private generation of sexually suggestive characters that appear childlike.

Meanwhile, Meta CEO Mark Zuckerberg continues to tout his company's AI capabilities. "We have a really exciting roadmap for this year with a unique vision focused on personalization," he said on an earnings call last month. "We believe that people don't all want to use the same AI—people want their AI to be personalized to their context, their interests, their personality, their culture, and how they think about the world."

But in a world where more and more people are turning to AI for companionship, Meta must weigh the risks of enabling such open-ended personalization.

From: Liz [REDACTED]@meta.com>

Sent: Tuesday, February 11, 2025 1:34 PM

To: Ahmad [REDACTED]@meta.com>; Connor Hayes [REDACTED]@meta.com>; Khushboo [REDACTED]@meta.com>; Joel Kaplan [REDACTED]@meta.com>; David Ginsberg [REDACTED]@meta.com>; Tucker Bounds [REDACTED]@meta.com>; Kevin Martin <[REDACTED]@meta.com>; Molly [REDACTED]@meta.com>; Josh Ginsberg [REDACTED]@meta.com>; Andrea Saul [REDACTED]@meta.com>; Ana [REDACTED]@meta.com>; Molly Cutler [REDACTED]@meta.com>; Guy Rosen [REDACTED]@meta.com>; Brian Rice [REDACTED]@meta.com>; Lindsay Elin [REDACTED]@meta.com>; Antigone Davis [REDACTED]@meta.com>; Jonathan D. Martin [REDACTED]@meta.com>; Nell McCarthy [REDACTED]@meta.com>; Adam Mosseri [REDACTED]@meta.com>; Greg Maurer [REDACTED]@meta.com>; Monika Bickert [REDACTED]@meta.com>; Scott Tucker [REDACTED]@meta.com>

Cc: Lena Pietsch [REDACTED]@meta.com>; Chris Norton [REDACTED]@meta.com>; Rachel Lieber [REDACTED]@meta.com>; AlComms [REDACTED]@meta.com>; Strategic Response Comms [REDACTED]@meta.com>; Katie [REDACTED]@meta.com>; Stephanie [REDACTED]@meta.com>; Sophie Vogel [REDACTED]@meta.com>; Faith [REDACTED]@meta.com>; Kate [REDACTED]@meta.com>; Paige [REDACTED]@meta.com>; Tyler [REDACTED]@meta.com>; Anthony [REDACTED]@meta.com>; Nicole [REDACTED]@meta.com>; Mel [REDACTED]@meta.com>

Subject: [FLAG]: Fast Company Story on AI Studio

Hi all,

Flagging that we're expecting a tough story from Fast Company on AI Studio this week. The story will focus on the types of AIs found in the "popular" tab of AI Home and AIs the reporter was able to create, specifically ones that were highly sexualized and some that appeared childlike. The reporter did not publish the AIs which the story will make clear, but they claim there were a lot of easy workarounds to create them.

We shared an on the record statement that speaks to our policies against child nudity and exploitation, and the enforcement and detection measures we have in place during the creation and publication of AIs in AI Studio.

Statement: "Our policies prohibit the publication of AIs that sexualize or exploit children. We have certain detection measures that work to prevent the creation of violating AIs, and published AIs are subject to the full extent of our detection and enforcement systems. We use sophisticated technology and reports from our community to identify and remove violating content."

We will flag as soon as it hits.

Thanks,
Liz