The Taylor Swift deepfake debacle was frustratingly preventable

Amanda Silberling

@asilbwrites / 2:37 pm PST • January 30, 2024

Image Credits: Kevin Winter / Getty Images

You know you've screwed up when you've simultaneously angered the White House, the TIME Person of the Year and pop culture's most rabid fanbase. That's what happened last week to X, the Elon Musk-owned platform formerly called Twitter, when AI-generated, pornographic deepfake images of Taylor Swift went viral.

One of the most widespread posts of the nonconsensual, explicit deepfakes was viewed more than 45 million times, with hundreds of thousands of likes. That doesn't even factor in all the accounts that reshared the images in separate posts — once an image has been circulated that widely, it's basically impossible to remove.

X lacks the infrastructure to identify abusive content quickly and at scale. Even in the Twitter days, this issue was difficult to remedy, but it's become much worse since Musk gutted so much of Twitter's staff, including the majority of its trust and safety teams. So, Taylor Swift's massive and passionate fanbase took matters into their own hands, flooding search results for queries like "taylor swift ai" and "taylor swift deepfake" to make it more difficult for users to find the abusive images. As the White House's press secretary called on Congress to do something, X simply banned the search term "taylor swift" for a few days. When users searched the musician's name, they would see a notice that an error had occurred.

This content moderation failure became a national news story, since Taylor Swift is Taylor Swift. But if social platforms can't protect one of the most famous women in the world, who can they protect?

"If you have what happened to Taylor Swift happen to you, as it's been happening to so many people, you're likely not going to have the same amount of support based on clout, which means you won't have access to these really important communities of care," Dr. Carolina Are, a fellow at Northumbria

University's Centre for Digital Citizens in the U.K., told TechCrunch. "And these communities of care are what most users are having to resort to in these situations, which really shows you the failure of content moderation."

Banning the search term "taylor swift" is like putting a piece of Scotch tape on a burst pipe. There are many obvious workarounds, like how TikTok users search for "seggs" instead of sex. The search block was something that X could implement to make it look like they're doing something, but it doesn't stop people from just searching "t swift" instead. Copia Institute and Techdirt founder Mike Masnick called the effort "a sledge hammer version of trust & safety."

"Platforms suck when it comes to giving women, non-binary people and queer people agency over their bodies, so they replicate offline systems of abuse and patriarchy," Are said. "If your moderation systems are incapable of reacting in a crisis, or if your moderation systems are incapable of reacting to users' needs when they're reporting that something is wrong, we have a problem."

So, what should X have done to prevent the Taylor Swift fiasco?

Are asks these questions as part of her research, and proposes that social platforms need a complete overhaul of how they handle content moderation. Recently, she conducted a series of roundtable discussions with 45 internet users from around the world who are impacted by censorship and abuse to issue recommendations to platforms about how to enact change.

One recommendation is for social media platforms to be more transparent with individual users about decisions regarding their account or their reports about other accounts.

“You have no access to a case record, even though platforms do have access to that material — they just don’t want to make it public,” Are said. “I think when it comes to abuse, people need a more personalized, contextual and speedy response that involves, if not face-to-face help, at least direct communication.”

X announced this week that it would hire 100 content moderators to work out of a new “Trust and Safety” center in Austin, Texas. But under Musk’s purview, the platform has not set a strong precedent for protecting marginalized users from abuse. It can also be challenging to take Musk at face value, as the mogul has a long track record of failing to deliver on his promises. When he first bought Twitter, Musk declared he would form a content moderation council before making major decisions. This did not happen.

In the case of AI-generated deepfakes, the onus is not just on social platforms. It’s also on the companies that create consumer-facing generative AI products.

According to an investigation by 404 Media, the abusive depictions of Swift came from a Telegram group devoted to creating nonconsensual, explicit deepfakes. The members of the group often use Microsoft Designer, which draws from OpenAI’s DALL-E 3 to generate images based on inputted prompts. In a loophole that Microsoft has since addressed, users could generate images

of celebrities by writing prompts like “taylor ‘singer’ swift” or “jennifer ‘actor’ aniston.”

A principal software engineering lead at Microsoft, Shane Jones, wrote a letter to the Washington state attorney general stating that he found vulnerabilities in DALL-E 3 in December, which made it possible to “bypass some of the guardrails that are designed to prevent the model from creating and distributing harmful images.”

Jones alerted Microsoft and OpenAI to the vulnerabilities, but after two weeks, he had received no indication that the issues were being addressed. So, he posted an open letter on LinkedIn to urge OpenAI to suspend the availability of DALL-E 3. Jones alerted Microsoft to his letter, but he was swiftly asked to take it down.

“We need to hold companies accountable for the safety of their products and their responsibility to disclose known risks to the public,” Jones wrote in his letter to the state attorney general. “Concerned employees, like myself, should not be intimidated into staying silent.”

OpenAI told TechCrunch that it immediately investigated Jones’ report and found that the technique he outlined did not bypass its safety systems.

“In the underlying DALL-E 3 model, we’ve worked to filter the most explicit content from its training data including graphic sexual and violent content, and have developed robust image classifiers that steer the model away from

generating harmful images," a spokesperson from OpenAI said. "We've also implemented additional safeguards for our products, ChatGPT and the DALL-E API – including declining requests that ask for a public figure by name."

OpenAI added that it uses external red teaming to test products for misuse. It's still not confirmed if Microsoft's program is responsible for the explicit Swift deepfakes, but the fact stands that as of last week, both journalists and bad actors on Telegram were able to use this software to generate images of celebrities.

Jones refutes OpenAI's claims. He told TechCrunch, "I am only now learning that OpenAI believes this vulnerability does not bypass their safeguards. This morning, I ran another test using the same prompts I reported in December and without exploiting the vulnerability, OpenAI's safeguards blocked the prompts on 100% of the tests. When testing with the vulnerability, the safeguards failed 78% of the time, which is a consistent failure rate with earlier tests. The vulnerability still exists."

As the world's most influential companies bet big on AI, platforms need to take a proactive approach to regulate abusive content — but even in an era when making celebrity deepfakes wasn't so easy, violative behavior easily evaded moderation.

"It really shows you that platforms are unreliable," Are said. "Marginalized communities have to trust their followers and fellow users more than the people

that are technically in charge of our safety online."

Updated, 1/30/24 at 10:30 PM ET, with comment from OpenAI

Updated, 1/31/24 at 6:10 PM ET, with additional comment from Shane Jones

> **Swift retaliation: Fans strike back after explicit deepfakes flood X**

> **Ahead of congressional hearing on child safety, X announces plans to hire 100 moderators in Austin**

Latest Stories

The bill that could ban TikTok passes in the House

Taylor Hatmaker

7:38 am PDT • March 13, 2024

Startups are hiring fewer workers, and paying out less in equity comp

Theresa Loconsolo

7:21 am PDT • March 13, 2024

Google Deepmind trains a video game-playing AI to be your co-op companion

Devin Coldewey

7:01 am PDT • March 13, 2024

TikTok ban: What's going on with the proposed bill in Congress

Taylor Hatmaker

6:40 am PDT • March 13, 2024

Furno bets that super-efficient, modular kilns will turn the cement industry upside down

Tim De Chant

6:31 am PDT • March 13, 2024

Blockchain startup Sei Labs creates an interesting solution to make Ethereum faster

Jacquelyn Melinek

6:00 am PDT • March 13, 2024