Deepfakes for all: Uncensored AI art model prompts ethics questions

Kyle Wiggers

@kyle_l_wiggers / 8:15 AM EDT • August 24, 2022

Comment

Image Credits: Bryce Durbin / TechCrunch

A new open source AI image generator capable of producing realistic pictures from any text prompt has seen stunningly swift uptake in its first week. Stability AI's Stable Diffusion, high fidelity but capable of generator services like Artbreeder, Pixelz.ai and more. But the model's unfiltered nature means not all the use has been completely above board.

For the most part, the use cases have been above board. For example, NovelAI has been experimenting with Stable Diffusion to produce art that can accompany the AI-generated stories created by users on its platform. Midjourney has launched a beta that taps Stable Diffusion for greater photorealism.

But Stable Diffusion has also been used for less savory purposes. On the infamous discussion board 4chan, where the model leaked early, several threads are dedicated to AI-generated art of nude celebrities and other forms of generated pornography.

Emad Mostaque, the CEO of Stability AI, called it "unfortunate" that the model leaked on 4chan and stressed that the company was working with "leading ethicists and technologies" on safety and other mechanisms around responsible release. One of these overall Stable Diffusion software package that attempts to detect and block offensive or undesirable images.

However, Safety Classifier — while on by default — can be disabled.

Stable Diffusion is very much new territory. Other AI art-generating systems, like OpenAI's DALL-E 2, have implemented strict filters for pornographic material. (The license for the open source Stable Diffusion prohibits certain applications, like exploiting minors, but the model itself isn't fettered on the technical level.) Moreover, many don't have the ability to create art of public figures, unlike Stable Diffusion. Those two capabilities could be risky when combined, allowing bad actors to create pornographic "deepfakes" that — worst-case scenario — might perpetuate abuse or implicate someone in a crime they didn't commit.

Women, unfortunately, are most likely by far to be the victims of this. A study carried out in 2019 revealed that, of women. That bodes poorly for the future of these AI systems, according to Ravit Dotan, VP of responsible AI at Mission Control.

“I worry about other effects of synthetic images of illegal content — that it will exacerbate the illegal behaviors that are portrayed,” Dotan told TechCrunch via email. “E.g., will synthetic child [exploitation] increase the creation of authentic child [exploitation]? Will it increase the number of pedophiles’ attacks?”

Montreal AI Ethics Institute principal researcher Abhishek Gupta shares this view. “We really need to think about the lifecycle of the AI system which includes post-deployment use and monitoring, and think about how we can envision controls that can minimize harms even in worst-case scenarios,” he said. “This is particularly true when a powerful capability [like Stable Diffusion] gets into the wild that can cause real trauma to those against whom such a system might be used, for example, by

Something of a preview played out over the past year when, at the advice of a nurse, a father took pictures of his young child’s swollen genital area and texted them to the nurse’s iPhone. The photo automatically backed up to Google Photos and was flagged by the company’s AI filters as child sexual abuse material, which resulted in the man’s account being disabled and an investigation by the San Francisco Police Department.

If a legitimate photo could trip such a detection system, experts like Dotan say, there’s no reason deepfakes generated by a system like Stable Diffusion couldn’t — and at scale.

“The AI systems that peofple create, even when they have the best intentions, can be used in harmful ways that they don’t anticipate and can’t prevent,” Dotan said. “I think that developers and researchers often underappreciated this point.”

Of course, the technology to create deepfakes has existed for some time, company Sensity found that hundreds of explicit deepfake videos featuring female celebrities were being uploaded to the world's biggest pornography websites every month; the report estimated the total number of deepfakes online at around 49,000, over 95% of which were porn. Actresses including Emma Watson, Natalie Portman, Billie Eilish and Taylor Swift have been the targets of deepfakes since AI-powered face-swapping tools entered the mainstream several years ago, and some, including Kristen Bell, have spoken out against what they view as sexual exploitation.

But Stable Diffusion represents a newer generation of systems that can create incredibly — if not perfectly — convincing fake images with minimal work by the user. It's also easy to install, requiring no more than a few setup files and a graphics card costing several hundred dollars on the high end. Work is underway on even more efficient versions of the system that can run on an M1 MacBook.

University of London, thinks the automation and the possibility to scale up customized image generation are the big differences with systems like Stable Diffusion — and main problems. “Most harmful imagery can already be produced with conventional methods but is manual and requires a lot of effort,” he said. “A model that can produce near-photorealistic footage may give way to personalized blackmail attacks on individuals.”

Berns fears that personal photos scraped from social media could be used to condition Stable Diffusion or any such model to generate targeted pornographic imagery or images depicting illegal acts. There’s certainly precedent. After reporting on the rape of an eight-year-old Kashmiri girl in 2018, Indian investigative journalist Rana Ayyub became the target of Indian nationalist trolls, some of whom created deepfake porn with her face on another person’s body. The deepfake was shared by the leader of the nationalist political party BJP, and the harassment Ayyub received as a

“Stable Diffusion offers enough customization to send out automated threats against individuals to either pay or risk having fake but potentially damaging footage being published,” Berns continued. “We already see people being extorted after their webcam was accessed remotely. That infiltration step might not be necessary anymore.”

With Stable Diffusion out in the wild and already being used to generate pornography — some non-consensual — it might become incumbent on image hosts to take action. TechCrunch reached out to one of the major adult content platforms, OnlyFans, who said that it would “continuously” update its technology to “address the latest threats to creator and fan safety, including deepfakes.”

“All content on OnlyFans is reviewed with state-of-the-art digital technologies and then manually reviewed by our trained human moderators to ensure that any person featured in the content is a verified spokesperson said via email. "Any content which we suspect may be a deepfake is deactivated."

A spokesperson for Patreon, which also allows adult content, noted that the company has a policy against deepfakes and disallows images that "repurpose celebrities' likenesses and place non-adult content into an adult context."

"Patreon constantly monitors emerging risks, like [AI-generated deepfakes]. Today, we do have policies in place that don't allow abusive behavior to real people and that disallows anything that could cause real world harm," the Patreon spokesperson continued in an email. "As technology or new potential risks emerge, we'll follow the process we have in place: working closely with creators to craft policies for Patreon, including what benefits are allowed and what kind of content is within guidelines."

DALL-E 2-like AI free, consequences be damned

DALL-E 2, OpenAI's powerful text-to-image AI system, can create photos in the style of cartoonists, 19th century daguerreotypists, stop-motion animators and more. But it has an important, artificial limitation: a filter that prevents it from creating images depicting public figures and content deemed too toxic. Now an open

If history is any indication, however, enforcement will likely be uneven — in part because few laws specifically protect against deepfaking as it relates to pornography. And even if the threat of legal action pulls some sites dedicated to objectionable AI-generated content under, there's nothing to prevent new ones from popping up.

In other words, Gupta says, it's a brave new world.

"Creative and malicious users can abuse the capabilities [of Stable Diffusion] to generate subjectively objectionable content at scale, using entire model — and then publish them in venues like 4chan to drive traffic and hack attention,” Gupta said. “There is a lot at stake when such capabilities escape out ‘into the wild’ where controls such as API rate limits, safety controls on the kinds of outputs returned from the system are no longer applicable.”

Editor’s note: An earlier version of this article included images depicting some of the celebrity deepfakes in question, but those have since been removed.

https://tcrn.ch/3RoueuT Copy

Tags

Abuse AI art generation

artificial intelligence deepfakes

image generation machine learning

open source Porn Stability AI

Ford, VW seeking buyer for Argo AI’s