Deepfake porn could be a growing problem amid AI race

April 16 2023, by Haleluya Hadero

Australian Noelle Martin poses for a photo Thursday, March 9, 2023, in New York. The 28-year-old found deepfake porn of herself 10 years ago when out of curiosity one day she used Google to search an image of herself. Credit: AP Photo/Andres Kudacki

Artificial intelligence imaging can be used to create art, try on clothes in

virtual fitting rooms or help design advertising campaigns.

But experts fear the darker side of the easily accessible tools could worsen something that primarily harms women: nonconsensual deepfake pornography.

Deepfakes are videos and images that have been digitally created or altered with artificial intelligence or machine learning. Porn created using the technology first began spreading across the internet several years ago when a Reddit user shared clips that placed the faces of female celebrities on the shoulders of porn actors.

Since then, deepfake creators have disseminated similar videos and images targeting online influencers, journalists and others with a public profile. Thousands of videos exist across a plethora of websites. And some have been offering users the opportunity to create their own images—essentially allowing anyone to turn whoever they wish into sexual fantasies without their consent, or use the technology to harm former partners.

The problem, experts say, grew as it became easier to make sophisticated and visually compelling deepfakes. And they say it could get worse with the development of generative AI tools that are trained on billions of images from the internet and spit out novel content using existing data.

"The reality is that the technology will continue to proliferate, will continue to develop and will continue to become sort of as easy as pushing the button," said Adam Dodge, the founder of EndTAB, a group that provides trainings on technology-enabled abuse. "And as long as that happens, people will undoubtedly ... continue to misuse that technology to harm others, primarily through online sexual violence, deepfake pornography and fake nude images."

Noelle Martin, of Perth, Australia, has experienced that reality. The 28-year-old found deepfake porn of herself 10 years ago when out of curiosity one day she used Google to search an image of herself. To this day, Martin says she doesn't know who created the fake images, or videos of her engaging in sexual intercourse that she would later find. She suspects someone likely took a picture posted on her social media page or elsewhere and doctored it into porn.

Horrified, Martin contacted different websites for a number of years in an effort to get the images taken down. Some didn't respond. Others took it down but she soon found it up again.

"You cannot win," Martin said. "This is something that is always going to be out there. It's just like it's forever ruined you."

The more she spoke out, she said, the more the problem escalated. Some people even told her the way she dressed and posted images on social media contributed to the harassment—essentially blaming her for the images instead of the creators.

Eventually, Martin turned her attention towards legislation, advocating for a national law in Australia that would fine companies 555,000 Australian dollars ($370,706) if they don't comply with removal notices for such content from online safety regulators.

But governing the internet is next to impossible when countries have their own laws for content that's sometimes made halfway around the world. Martin, currently an attorney and legal researcher at the University of Western Australia, says she believes the problem has to be controlled through some sort of global solution.

In the meantime, some AI models say they're already curbing access to explicit images.

OpenAI says it removed explicit content from data used to train the image generating tool DALL-E, which limits the ability of users to create those types of images. The company also filters requests and says it blocks users from creating AI images of celebrities and prominent politicians. Midjourney, another model, blocks the use of certain keywords and encourages users to flag problematic images to moderators.

Meanwhile, the startup Stability AI rolled out an update in November that removes the ability to create explicit images using its image generator Stable Diffusion. Those changes came following reports that some users were creating celebrity inspired nude pictures using the technology.

Stability AI spokesperson Motez Bishara said the filter uses a combination of keywords and other techniques like image recognition to detect nudity and returns a blurred image. But it's possible for users to manipulate the software and generate what they want since the company releases its code to the public. Bishara said Stability AI's license "extends to third-party applications built on Stable Diffusion" and strictly prohibits "any misuse for illegal or immoral purposes."

Some social media companies have also been tightening up their rules to better protect their platforms against harmful materials.

TikTok said last month all deepfakes or manipulated content that show realistic scenes must be labeled to indicate they're fake or altered in some way, and that deepfakes of private figures and young people are no longer allowed. Previously, the company had barred sexually explicit content and deepfakes that mislead viewers about real-world events and cause harm.

The gaming platform Twitch also recently updated its policies around

explicit deepfake images after a popular streamer named Atrioc was discovered to have a deepfake porn website open on his browser during a livestream in late January. The site featured phony images of fellow Twitch streamers.

Twitch already prohibited explicit deepfakes, but now showing a glimpse of such content—even if it's intended to express outrage—"will be removed and will result in an enforcement," the company wrote in a blog post. And intentionally promoting, creating or sharing the material is grounds for an instant ban.

Other companies have also tried to ban deepfakes from their platforms, but keeping them off requires diligence.

Apple and Google said recently they removed an app from their app stores that was running sexually suggestive deepfake videos of actresses to market the product. Research into deepfake porn is not prevalent, but one report released in 2019 by the AI firm DeepTrace Labs found it was almost entirely weaponized against women and the most targeted individuals were western actresses, followed by South Korean K-pop singers.

The same app removed by Google and Apple had run ads on Meta's platform, which includes Facebook, Instagram and Messenger. Meta spokesperson Dani Lever said in a statement the company's policy restricts both AI-generated and non-AI adult content and it has restricted the app's page from advertising on its platforms.

In February, Meta, as well as adult sites like OnlyFans and Pornhub, began participating in an online tool, called Take It Down, that allows teens to report explicit images and videos of themselves from the internet. The reporting site works for regular images, and AI-generated content—which has become a growing concern for child safety groups.

"When people ask our senior leadership what are the boulders coming down the hill that we're worried about? The first is end-to-end encryption and what that means for child protection. And then second is AI and specifically deepfakes," said Gavin Portnoy, a spokesperson for the National Center for Missing and Exploited Children, which operates the Take It Down tool.

"We have not ... been able to formulate a direct response yet to it," Portnoy said.