Keeping it real: How to spot a deepfake

February 10 2024, by Madeleine Clarke

Created by our researchers using Midjourney, these diffusion model deepfakes illustrate the increasing realism and sophistication of the technology. Credit: CSIRO

In a world where you can create a virtual clone of a person in a matter of minutes, how do we know what's real? It may sound like dystopian science fiction, but deepfakes are a reality causing serious social, financial and personal harm.

Deepfakes are synthetic media generated using artificial intelligence (AI), including images, videos, audio and even text. Most commonly, deepfakes are videos or images of people that have been digitally manipulated by cyber attackers to mislead. Deepfakes can depict people saying or doing something they never actually did.

Due to recent advances in the AI used to create deepfakes, and the proliferation of cheap and easy-to-use deepfake generators, you've likely noticed more realistic deepfakes popping up on your social media feeds. Or maybe, more concerningly… you haven't noticed.

How are deepfakes made?

The "deep" in deepfake comes from deep learning, the kind of AI used to create them. Deep learning teaches computers to process data and make predictions in a way that is inspired by the human brain.

Until recently, many deepfakes were created using a type of model called a Generative Adversarial Network (or GAN). A GAN is trained by pitting two neural networks against each other: one generates content and the other evaluates it, which creates increasingly realistic outputs through competition.

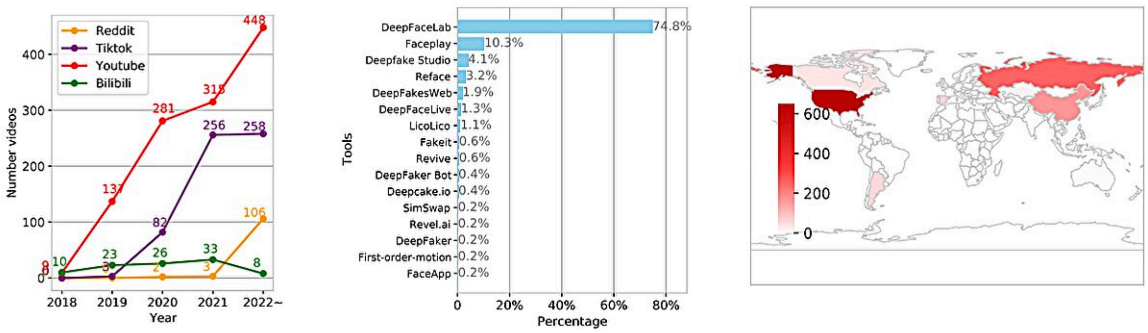
In deepfake videos created by GAN models, manipulation is applied to specific parts of the footage, for example the mouth. This results in fakes like this one where Twiggy Forest, Gina Reinhardt and Dick Smith's likenesses were used to scam Aussies to invest in a bogus scheme.

Since the boom in generative AI in 2023, there has been a rise in diffusion model deepfakes. Diffusion models take us beyond the realm of pasting a celebrity's face onto the body of an actor, or making the mouth say new words. They allow a deepfake to be created from scratch, without editing original content.

Our cybersecurity experts are researching the rapid changes in the deepfake landscape. Dr. Sharif Abuadbba said as well as increasing realism, the generative AI boom has led to cheaper and more accessible deepfake creation.

"Just a year or so ago this technology was only accessible to skilled hackers and experts. Now, anyone with a phone or computer can make a deepfake. It's as easy as going to the app store, downloading one of many apps for free or a small subscription fee, and you're ready to go," Sharif said.

"Scarily, it's now almost just as easy to create a deepfake that targets a specific person. Novice attackers can access a generator and create one with no prior knowledge in as little as a few seconds. Skilled hackers can create highly realistic deepfakes in just a few hours to a day."



(a) Growth across platforms

(b) Proportion of deepfake applications

(c) Geographic distribution of publisher

Insights on the platform and country of origin, and app used to create deepfakes from our analysis of more than 2000 deepfakes in the wild. Credit: CSIRO

Why do people make deepfakes?

Sadly, the overwhelming majority of deepfakes target women and are used in pornography. However, other uses of deepfakes are growing, many of them malicious. They've been used for election tampering, identity fraud, scam attempts, and to spread fake news. Last year, the stock market tumbled in response to a realistic deepfake photo showing the Pentagon on fire.

Our experts recently collaborated with researchers from Sungkyunkwan University in South Korea to collect and study the largest and most diverse dataset of real deepfakes in the world. It's made up of 2,000 deepfakes from 21 countries in English, Russian, Mandarin, and Korean languages. The deepfakes were sourced from Youtube, TikTok, Reddit, and Chinese video sharing platform Bilibili.

Their analysis found deepfakes in the entertainment category—think Robert Pattinson dancing—doubled every year from 2019 to 2021. There has also been a substantial growth in political and fraud deepfakes.

Sharif said the use of deepfakes is becoming more common and emerging as a serious cybersecurity threat.

"Deepfakes are getting so advanced in their realism that they have the potential to target facial recognition systems that are increasingly being used to secure your online accounts like banking," he said.

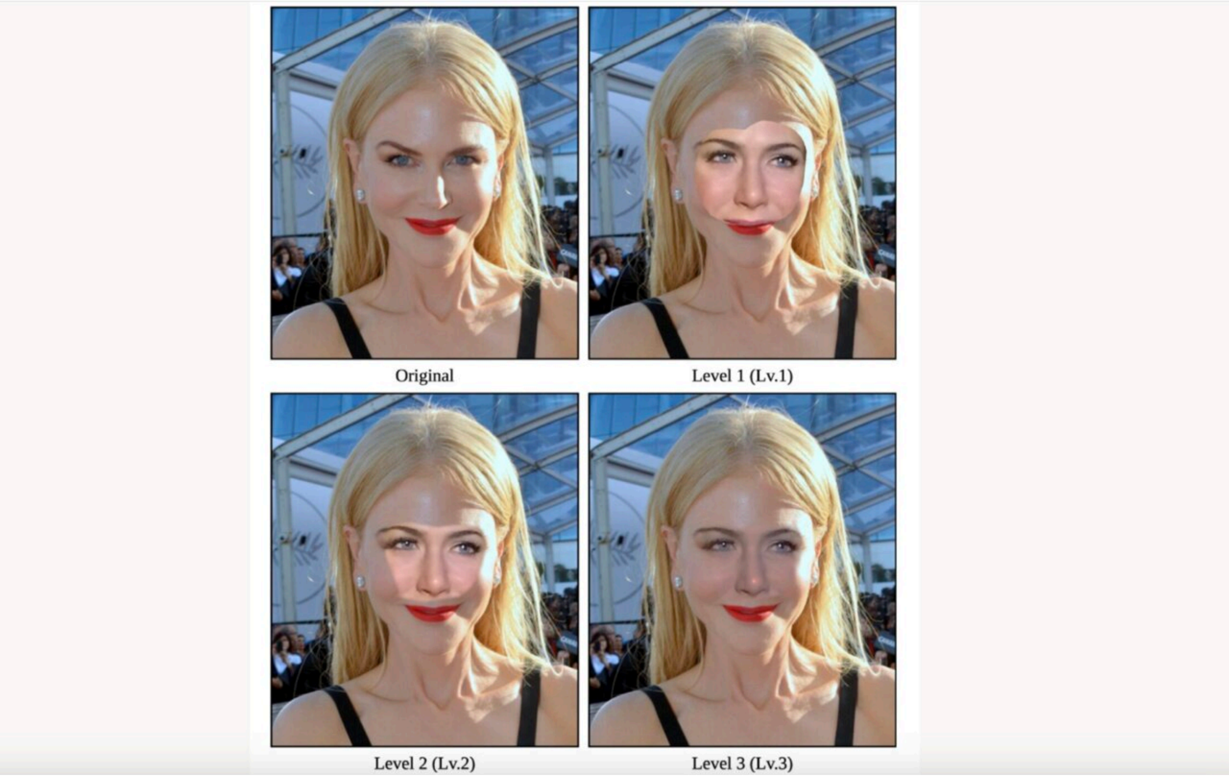
Gartner have predicted deepfakes could shatter trust in biometric authentication for 30% of companies by 2026. There is also rising concern about the ability to create voice clones to hack voice protected online accounts or things like google assistants.

How to spot a deepfake

Given the explosion of new deepfakes on our social feeds and their potential to cause real harm, it's important to know how to spot them.

Our expert Dr. Kristen Moore said there are different features to look out for that can expose the lie.

"If it's a video, you can check if the audio is properly synced to the lip movement. Do the words match the mouth? Other things to check for are unnatural blinking or flickering around the eyes, odd lighting or shadows, and facial expressions that don't match the emotional tone of the speech," Kristen said.



A deepfake of Jennifer Anniston using a source image of Nicole Kidman created using the face swap method. Credit: Source: Lee, Tariq, Shin & Woo, 2021

"Deepfake images made from diffusion models can typically be spotted by looking for asymmetries, like earrings that are different on each side, or one eye being slightly bigger than the other. Another thing they currently really struggle with is hands—so check for the right number of fingers, and the size and realism of the hand.

"For deepfakes that have used the face swap method, you can sometimes see the point where they've blended the face onto the original forehead. Sometimes there's a textural or color difference, or maybe the hairline just looks a bit off."

However, she also cautions that deepfakes may soon become undetectable to anyone but the most highly trained experts.

"The jump in capability following the introduction of the GPT model in the text space has just blown everyone's minds. Generative AI went from being not particularly sophisticated to something many people are now using every day… it's phenomenal," Kristen said.

"As the generators are improving so quickly, we encourage people to be skeptical about the content they see online. Fact-check content by comparing the media in question with that from other trusted sources, and always seek out the original source of the content."

How to protect yourself from deepfakes

Our cybersecurity researchers including Sharif, Kristen, and Dr. Shahroz Tariq at Data61 are working on digital methods to combat deepfake attacks, including watermarking authentic content and improving the accuracy of AI-powered automatic deepfake detectors. But there is some way to go before digital methods can reliably detect real deepfakes.

With this in mind, it's vital that we take steps to protect ourselves.

Shahroz said preventing the creation of deepfakes is technically impossible to achieve for celebrities or public figures given the sheer volume of videos and images of them online. But there are steps the public can take to avoid being deepfaked.

"You can protect yourself to some extent by making your social profiles private. This way only your friends and followers can view your content," Shahroz said.

"Not only does this limit the availability of your images online in general, but it significantly increases the potential of finding the attacker if someone does create a deepfake of you."

Sharif said organizations in industries who collect individuals' identifying data are also particularly vulnerable and should be proactive to get ahead of the threat.

"If you're in industries like news and entertainment, recruitment (where we're seeing a rise of video job applications), social media, banking, and institutions using facial recognition… deepfakes are either coming to you or they're already there," Sharif said.

"We're keen to work with industry on this problem and encourage anyone interested in collaborating with us to get in touch."

Provided by CSIRO

Citation: Keeping it real: How to spot a deepfake (2024, February 10) retrieved 27 April 2024 from https://techxplore.com/news/2024-02-real-deepfake.html