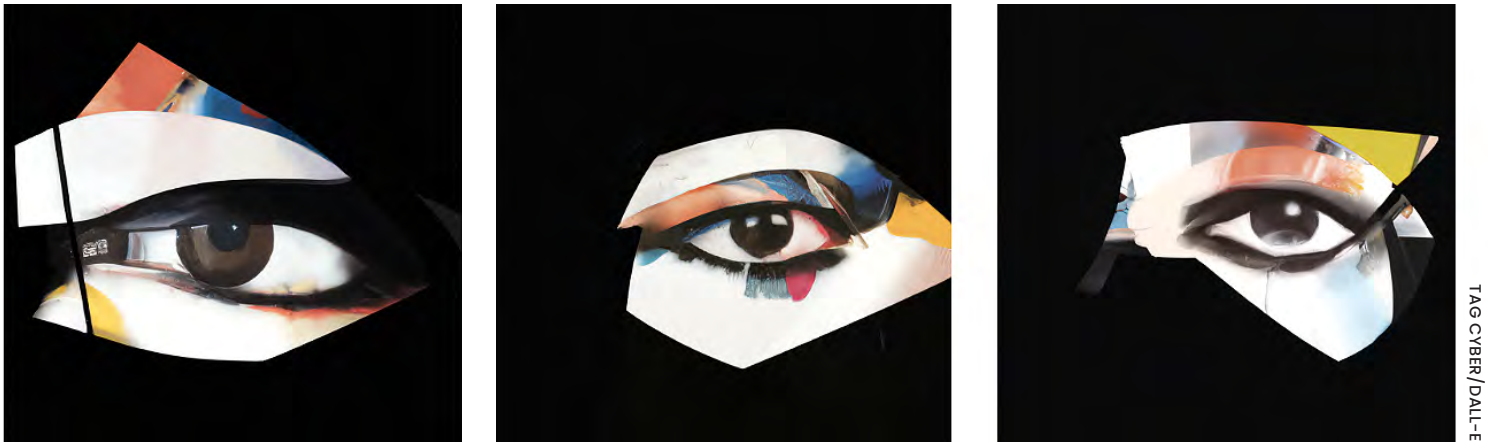
THERE'S MORE TO DEEPFAKES THAN MEETS THE EYE

DAVID HECHLER



TAG CYBER/DALL-E

What do you think of when you hear the word "deepfakes"? A video featuring Tom Cruise saying and doing silly things? A series of photographs with a face morphing from male to female? A clip of Kim Jong-un in which he addresses the American public? A guy who used to post on Reddit?

Some of you may be hearing (or seeing) that word for the first time. Others know a lot about it. They know that it got its name from a guy who used it on Reddit. And they've seen lots of Tom Cruise memes. They understand that, even though many people think immediately of videos, there are also deepfake audios. And I didn't even mention those, or pornography, in the paragraph above. So you see, there's a wider variety of deepfakes than some people realize.

Let's start with the basics. As the term is understood today, it combines **"deep learning"**—a kind of machine learning—and "fakes." What you're seeing or hearing is not the real thing: Deefakes are built from manipulated sounds and/or images. But the motives behind the manipulation are not all the same. That's why they shouldn't all be lumped together.

THEY'RE NOT ALL BAD

Deepfakes have a bad reputation. The ones that get the most attention are those in which the content manipulators do not ask the people featured in the fakes for permission to use their voices or images, and their motives may be malicious or indifferent to how the individuals affected may feel. But lots of deepfakes are created for amusement and seem harmless. They may be satire or parody. Others are designed to make a serious political point. And many harbor no intent to deceive.

Kim Jong-un deepfake

REPRESENTUS

In fact, some deepfakes announce themselves as fakes. For instance, the **Kim Jong-un clip**, above, was created by the nonpartisan, nonprofit **RepresentUs** as a public service ad. The North Korean leader, seated at a desk and clad in a Mao jacket, calmly warns American voters that he doesn't have to work to destroy their country. He points to their partisan divisions and ferocious fights over elections. "It's not hard for democracy to collapse. All you have to do," he says, pausing to crack a smile, "is nothing." The film ends with these words on the screen: "This footage is not real, but the threat is."

Another public service spot used a **deepfake of Joaquin Oliver**, a Stoneman Douglas High School student who was killed in the Parkland, Florida, shooting. His parents introduced him by explaining in a video that he'd been gone for two years and had missed his first opportunity to vote in an election. Now artificial intelligence has allowed him to speak again. The deepfake video of their son follows, and he offers an impassioned plea for people to vote "because nothing's changed, people are still getting killed by guns." He urges them to vote "because I can't."

The many deepfakes of **Tom Cruise** make lighthearted fun of the actor, but in recent years actors have benefited from this new technology. When a documentary about the career of Val Kilmer was being filmed, the actor was not able to sit for an interview because an operation to treat his throat cancer had left his voice badly damaged. But a company called Sonatic has been able to recreate his voice in a way that has **extended** his acting career.

Then there's Bruce Willis, whose health problems led him to retire from acting. But he recently **made a deal** to allow a company called Deepcake (that's not a typo) to map his face onto the body of another actor for a **commercial**. Though there was some disagreement about the circumstances, the message Deepcake was

Joaquin Oliver deepfake

CHANGE THE REF

announcing was clear. As was the company's aim to launch a new industry. Actors who can no longer act, the company seemed to be saying, or actors who have a commitment to perform that conflicts with another opportunity elsewhere, can now digitally clone themselves by authorizing deepfakes.

GRAY AREAS

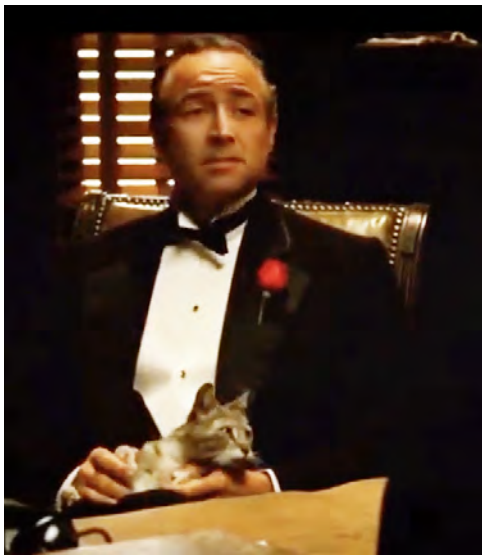
A director's failure to alert viewers that a voice was deepfaked in a recent documentary stirred controversy.

Some uses of deepfakes have been criticized on ethical grounds for failing to inform the audience. A noteworthy example involved a documentary about Anthony Bourdain that was filmed after he committed suicide. The director had access to thousands of hours of video and audio from his subject's popular food and travel television shows. But in three instances the director wanted to introduce sentences that Bourdain had written but had not recorded. So he decided to use deepfaked audio of Bourdain's voice.

When director Morgan Neville first acknowledged what he'd done, **several critics were aghast**—both that he'd done it and hadn't disclosed it in the film. I can't help but think that it won't be long before people simply accept such things, now that this is an option. I can imagine a far greater uproar had Neville inserted Bourdain deepfaked on video, but this, too, is easy to do. It seems bound to happen. And my guess is that it won't take long before the novelty, and ethical qualms, wear off.

By contrast, there was no need to issue a disclosure when Carrie Fisher and Peter Cushing made deepfaked appearances in "Rogue One: A Star Wars Story." They'd both been gone for years, of course. And one can be sure the use of their images was authorized. Somehow it seemed quite natural, given that this was a science fiction movie, after all. Now the **question** seems to be whether the Star Wars franchise will bring back Fisher, Mark Hamill and Harrison Ford for a deepfaked reunion—deepfaked to make them all youthful again, even though two are still alive. The money seems to say yes, and you can be sure that ethics won't stand in the way.

THE DARK SIDE



Nicolas Cage as Marlon Brando deepfake

As I noted earlier, the deepfakes that get the most attention are controversial. Obvious examples are the ones created by the Reddit user whose handle gave the concept its name. In late 2017, he began posting on Reddit pornographic videos in which the women's faces had been replaced by those of well-known actresses and other celebrities. As the popularity of his postings grew, he started a so-called Subreddit called deepfakes in which other registered users (known as Redditors) shared their own creations. In addition to pornography, Redditors posted deepfakes of other kinds of entertainment. A particularly popular series which became a genre unto itself offered deepfakes of **Nicolas Cage**. These were often compilations of brief movie clips in which Cage's face was swapped into the bodies of well-known actors and actresses ranging from Marlon Brando in a scene from "The Godfather," to Julie Andrews walking in the hills above Salzburg singing: "The hills are alive with the sound of music." Nothing dark or gray there. Unlike the hard-core content it was paired with, these were just silly.

The Deepfakes Subreddit was eventually shut down, and it wasn't because of the Cage videos. The network **banned** the Subreddit for violating its content policy, "specifically our policy against involuntary pornography," the announcement said. Deepfake pornography is still widely available elsewhere, of course. By at least one measure, it completely dominates the field. In 2019, an Amsterdam-based organization called Deeptrace issued a **report** that found that 96% of all deepfake videos online were pornographic.

Political deepfakes can pose grave dangers if they fool the public, but they're more easily defused because they're out in the open.

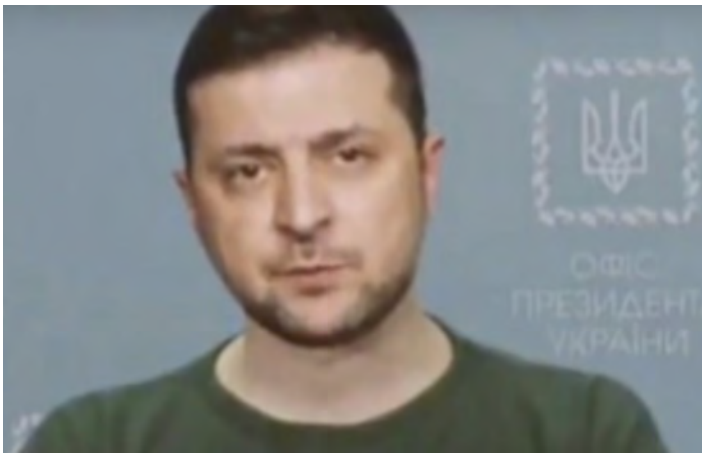
To put the Subreddit takedown in context, the unauthorized posting of pornographic images of women by men had been a serious problem since at least 2010. (These earlier postings did not involve deepfakes, but they paved the way for the Deepfakes Subreddit.) It was 2010 when Hunter Moore, from Woodland, California, started isanyoneup.com, the internet's best known "**revenge porn**" website. Moore encouraged people to submit real sexually explicit photographs of women without their consent, which he then posted on the site. They were often supplied by men who bore a grudge. California passed a **law** in 2013 making it crime to post this material knowing that it would cause the women emotional distress, and two years later **Moore** pleaded guilty and was sent to prison. In 2014, the "**Celebgate**" scandal broke in which at least five men hacked into the computers of more than 200 celebrities, including actresses Jennifer Lawrence and Mary Elizabeth Winstead, to steal nude photographs and other private material.

In the years that followed, technology made it easy for anyone to create deepfakes. By 2018, anyone could create them using software programs that were readily available. A short time later, celebrity deepfake videos were easy to create from a mobile phone.

PLAYING FOR HIGHER STAKES

Some of the most dangerous deepfakes have been ones that have targeted political leaders. The danger was in the potential consequences if they had been believed. During the U.S. presidential campaign in 2020, some videos promoted by the Trump campaign appeared to show Joe Biden as old, tired, confused and out of touch, but they **were actually deepfakes.**

Nearly two years later, Russia was engaged in a different kind of campaign. Three weeks after the country invaded Ukraine, a **deepfake of Ukraine President Volodymyr Zelensky** was broadcast showing him addressing his soldiers and instructing them to lay down their arms. The video was promoted by Russian social media along with posts on Facebook, Twitter and YouTube. In both instances, the targets quickly called out the fakes and they were removed from wide distribution. In Ukraine, the government had even warned its citizens in advance to expect Russia to engage in this kind of subterfuge.



Volodymy Zelensky deepfake

As serious as those incidents were, in one important respect they were easier to defuse than many other deepfakes for one simple reason: They were out in the open. That was the whole point. They were designed to influence public opinion. But that

also meant that they were closely scrutinized by journalists and experts of all stripes. It didn't take long to identify what they really were.

By contrast, criminals thrive on stealth. They often use deepfakes to try to trick businesses into wiring them funds, or they extort money by threatening to expose the image of a CEO in a compromising position. And companies are often reluctant to reveal anything about these episodes—whether they succeeded or failed, whether the images were genuine or phony—for fear of tarnishing their reputations. So it can be hard to know how big a threat deepfakes represent.

One indication that it's growing can be found in VMware's annual Global Incident Response Threat Report. In June 2022, it surveyed 125 cybersecurity and incident response professionals and found a 13% uptick in deepfakes year over year. And 66% of respondents had seen them during the previous 12 months, with email cited by 78% as the most common delivery method.

HELP NOT WANTED

This technology is new enough that innovations seem to pop up regularly. Here's a new twist. Now that so much work is conducted from remote locations far from traditional offices, it's no longer unusual for job interviews to be conducted remotely, and for employees to work for years for bosses they haven't met and may never meet. So perhaps it shouldn't be shocking that some companies have found they've hired not the fine young man or woman they thought they had, but a deepfake instead.

Last June, the FBI issued an **alert** that warned companies about deepfake job candidates. Complaints along these lines have been growing, the bureau noted. Rick McElroy, principal cybersecurity strategist at VMware, said it shouldn't be surprising. As companies have improved their security, criminals looked for other ways to break in. "Organizations have spent an inordinate amount of money on these controls," he said. "Manipulation of the human is the easiest way—it's the fast forward button."

Humans have even supplied the raw materials the criminals use to create deepfakes. We give them up ourselves when we post photos, videos and audio files on websites and on social media. And the ability of technology to turn stolen identities into deepfakes is improving rapidly. It isn't flawless, McElroy said. The FBI alert noted that audio and video are sometimes imperfectly synched, and that can help companies detect deepfakes. But in the hands of skillful criminals, it's often good enough.

For the criminals, there are real advantages in using this approach, McElroy continued. Human imposters might succeed in securing the same jobs, but they would be hard-pressed to apply for positions at companies around the country or around the world. Deepfakes can scale. And once they obtain employment, they can look for opportunities to steal money if their handlers are criminals, or engage in espionage if their owners are nation-states. (Or do both.)

What strikes me as particularly unsettling is that if you hire and eventually uncover the true "identities" of **deepfake employees**, you may still be left wondering who created them and who they really worked for.

Now that we've explored the wide range of deepfakes—from light entertainment to those that may be most important to consider, but also most unpleasant—this might be a good time to click on one of those "Tom Cruise" videos that you'll have no trouble locating on the 'net. I find they have a welcome calming effect.

(Reprinted from the TAG Cyber Security Annual, 1st Quarter 2023.)