Video and Audio Deepfakes: What Lawyers Need to Know

by Sharon D. Nelson, Esq., and John W. Simek

If some nefarious person has decent photos of your face, you too (like so many unfortunate Hollywood celebrities) could appear to be the star of a pornographic video. If someone has recordings of your voice (from your website videos, CLEs you have presented, speeches you've given, etc.), they can do a remarkably good job of simulating your spoken words and, just as an example, call your office manager and authorize a wire transfer – something the office manager may be willing to do because of "recognizing" your voice.

Unnerving? Yes, but it is the reality of today. And if you don't believe how "white hot" deepfakes are, just put a Google alert on that word and you'll be amazed at the volume of daily results.

Political and Legal Implications

We have already seen deepfakes used in the political area (the "drunk" Nancy Pelosi deepfake, a reference to which was tweeted by the president), and many commentators worry that deepfake videos will ramp up for the 2020 election. Some of them, including the Pelosi video, are referred to as "cheapfakes" because they are so poorly done (basically running the video at 75 percent speed to simulate drunkenness), but that really doesn't matter if large numbers of voters believe it's real. And the days when you could tell a deepfake video by the fact that the person didn't blink are rapidly vanishing as the algorithms have gotten smarter.

In August 2019, the Democratic National Committee wanted to demonstrate the potential threat to the 2020 election posed by deepfake videos so it showed, at the 2019 Def Con conference, a video of DNC Chair Tom Perez. The audience was told he was unable to come but would appear via Skype. Perez came on screen and apologized for not being in attendance—except that he had said no such thing. It was a deepfake.

Another deepfake video surfaced of Facebook CEO Mark Zuckerberg in June 2019, with him supposedly saying: "Imagine this for a second. One man, with total control of billions of people's stolen data. All their secrets, their lives, their futures. I owe it all to Spectre. Spectre showed me that whoever controls the data controls the future." It was hardly a credible fake, but since he appeared to be talking to CBS, CBS asked that Facebook remove the video, complaining about the unauthorized use of its trademark.

The COVID-19 pandemic has even provided a platform for deepfakes. In April 2020, a Belgium political group released a deepfake video showing the prime minister of Belgium calling for forceful climate control change due to the linkage of COVID-19 and environmental damage.

Just the possibility of a deepfake can cause doubts and questioning of the authenticity whether the video is fabricated or real. A real-world example comes from a small central Africa country in late 2018 called Gabon. Back then, the president of Gabon, Ali Bongo, failed to make a public appearance for months. Rumors began circulating that he was in poor health or even dead. The administration announced that Bongo would give a televised address on New Year's Day hoping to put the rumors to bed and reestablish faith in the government. The video address didn't go well as Bongo appeared to be unnatural with suspect facial mannerisms and unnatural speech patterns. The political opponents immediately claimed that the

video was a deepfake and the conspiracy quickly spread on social media sparking an unsuccessful coup by the military. Experts still can't definitely say if the video was authentic or not, but Bongo has since appeared in public. As USC professor Hao Li said, "People are already using the fact that deepfakes exist to discredit genuine video evidence. Even though there's footage of you doing or saying something, you can say it was a deepfake and it's very hard to prove otherwise."

Deepfakes are capable of influencing elections and perhaps the rule of law, which should certainly compel the attention of lawyers, especially since many lawyers regard the rule of law as already under fire.

Legislation has been introduced in Congress to do something about deepfakes to prevent an impact on our elections. It has gone nowhere. The First Amendment is often cited as an obstacle to legislation, as is the fair use provision of copyright law, existing state privacy, extortion and defamation laws, and the Digital Millennium Copyright Act, all for different reasons.

The Malicious Deep Fake Prohibition Act, introduced in Congress, would make it a federal crime to create a deepfake when doing so would facilitate illegal conduct. It was not well received. The DEEPFAKES Accountability Act requires mandatory watermarks and clear labeling on all deepfakes (oh sure, the bad guys will respect that law!). It contains a very broad definition of deepfakes, which almost certainly would guarantee that it would face a constitutional challenge. In short, we haven't gotten close to figuring out how to deal with deepfakes via legislation.

And yet, according to a June 2019 Pew Research Center survey, nearly two-thirds of Americans view altered videos and images as problematic and think something should be done to stop them. According to the survey, "Roughly three-quarters of U.S. adults (77 percent) say steps should be taken to restrict altered images and videos that are intended to mislead." The survey indicated that the majority of Republicans and Democrats believe that.

What Is a Deepfake Video?

No worries, we'll get to deepfake audios later. The audios are a new phenomenon in the toolbox of criminals, but deepfake videos have become—quickly—terrifyingly mainstream. Remember the "old" days of video manipulation when we were all amazed, watching Forrest Gump as he met President Kennedy 31 years after the president's assassination? Ah, the days of innocence!

We are writing here for lawyers, so we are not going into the weeds of how true deepfakes are produced. It is a remarkably complex process when well done. But we can give you a 10,000-foot picture of what deepfakes are and how, in vastly simplified terms, they are created.

We actually like the Wikipedia definition of deepfake: *"a portmanteau of 'deep learning' and 'fake' is a technique for human image synthesis based on artificial intelligence. It is used to combine and superimpose existing images and videos onto source images or videos using a machine learning technique known as generative adversarial network. The phrase 'deepfake' was coined in 2017. Because of these capabilities, deepfakes have been used to create fake celebrity pornographic videos or revenge porn. Deepfakes can also be used to create fake news and malicious hoaxes."*

If you thought Photoshop could do bad things, think of deepfakes as Photoshop on steroids!

Deepfake video is created by using two competing (and yet collaborative) AI systems—a generator and a discriminator. The generator makes the fake video and then the discriminator examines it and determines whether the clip is fake. If the discriminator correctly identifies a fake, the generator learns to avoid doing the same thing in the clip it creates next.

The generator and discriminator form something called a generative adversarial network (GAN). The first step in establishing a GAN is to identify the desired output and create a training dataset for the generator. Once the generator begins creating an acceptable level of output, video clips can be fed to the discriminator. As the generator gets better at creating fake video clips, the discriminator gets better at spotting them. Conversely, as the discriminator gets better at spotting fake video, the generator gets better at creating them. The discriminator and generator work iteratively against each other. Over time, the success rate of the discriminator drops towards 50%. Basically, the discriminator is randomly guessing as to whether the video is real or not. If thinking about that makes your head spin, you are not alone.

Yes, it is technical and not especially easy to understand, though there are (of course) Ikea-like DIY toolkits that will do most of the advanced work for you. It is particularly disturbing that free deepfake generators are available in the open-source community where there are no ethical rules enforced.

Want to try your hand at a video deepfake? Download and install the Reface app (previously known as Doublicat) on your iPhone or Android device. Reface uses AI technology to let users "Create hyper-realistic face swap videos & gifs with just one selfie."

Hany Farid on Deepfakes

Hany Farid, a professor who is often called the father of digital image forensics, is a special hero to the authors. His incredible work on detecting altered images has inspired us for years. Here is what Farid says about deepfakes:

"The technology to create sophisticated fakes is growing rapidly. Every three to six months, we see dramatic improvements in the quality and sophistication of the fakes. And, these tools are becoming increasingly easier to use by the average person. Deepfakes (image, audio and visual) can have broad implications ranging from nonconsensual pornography to disrupting democratic elections, sowing civil discord and violence, and fraud. And more generally, if we eventually live in a world where anything can be faked, how will we trust anything that we read or see online?"

That's a darn good question for which, at the moment, no one has a satisfactory answer.

Farid doesn't believe, and we agree, that we can truly control the deepfakes situation. To exercise any degree of control, according to him, "will require a multifaceted response from developing better detection technology to better policies on the large social media platforms, a better educated public and potentially legislation."

New photo and video verification platforms like Truepic (full disclosure: Farid is an advisor to the company) use blockchain technology to create and store digital signatures for authentically shot videos as they are being recorded, which makes them easier to verify later. But today, that is the exception and not the reality.

By the way, sharing deepfake revenge porn is now a crime in Virginia (effective July 1, 2019), the first state to make it a crime. We hope many more will follow. How do we combat the spread of $50 apps like DeepNude (thankfully defunct, but there will be others), which could undress women in a single click? DeepNude was trained on more than 10,000 images of nude women and would provide the undressed woman within 30 seconds—and of course the image could be shared to smear reputations (sending it to the woman's employer or friends and family) or to post online as revenge porn.

How Do We Tell If a Video Is a Deepfake?

Researchers can now examine videos for signs it's a fake, such as shadows and blinking patterns. In June 2019, a new paper from several digital forensics experts outlined a more foolproof approach that relies on training a detection algorithm to recognize the face and head movements of a particular person, thereby showing when that person's face has been imposed onto the head and body of someone else. The drawback is that this approach only works when the system has been trained to recognize particular people, but it might at least keep presidential candidates safe from attack.

Another problem is that technological advances are happening at lightning speed. Researchers published a study showing that irregularities with the eyes blinking could be used to identify a deepfake video. Unfortunately, several months later, the generators learned how to correct for the blinking imperfection putting detection back several steps.

In August 2019, the Defense Advanced Research Projects Agency (DARPA) announced a new initiative intended to curtail malicious deepfakes, which we all have seen spreading rapidly. Using the Semantic Forensics program, or SemaFor, researchers will train computers to spot deepfakes by using common sense. Don't ask us how that works. In any case, we fear that DARPA may be a bit late to the game as the 2020 election approaches.

The internet is rife with stories of how, when audio is a part of a deepfake video, our new best friends might be mice. There has been some fascinating new work at the University of Oregon's Institute of Neuroscience. There, they have a research team training mice to understand differences within human speech. So who knew that mice had that capability? While most people don't likely know what phenomes are, they are small sounds we make which make one word discernible from another. The mice were able to correctly identify sounds as much as 80% of the time. And since they were given treats when they were correct, they were no doubt listening closely. Imperfect? Sure, but another weapon in the arsenal we are developing to detect deepfakes.

Progress has been impressive. Still, the deepfakes are getting better all the time. Are we ready for a barrage of deepfake videos before the 2020 election? The almost unanimous answer of the experts is "no."

To return to Farid's colorful expressions of futility, "We are outgunned. The number of people working on the video-synthesis side, as opposed to the detector side, is 100 to 1." Those are daunting odds.

Audio Deepfakes

As if deepfake videos weren't driving us crazy trying to discern the real from the unreal, now voice-swapping has begun to be used in artificial intelligence cyberattacks on business, allowing attackers to gain access to corporate networks and persuade employees to authorize a money transfer.

Business email compromise (BEC) attacks have become woefully common. Generally, they start with a phishing email to get into an enterprise network and look at the payment systems. They are looking for the employees authorized to wire funds and the entities that they usually wire funds to.

It is a theatrical game, as they emulate the executive, scaring or persuading the employee. This is really ramping up the success of the acts as using the phone to impersonate an executive is a powerful tool.

As with video deepfakes, the AI has GANs that constantly teach and train each other, perfecting the outcome.

Also emulating the fake videos, the bad guys come up with a convincing voice model by giving training data to the algorithm. The data might come from speeches, presentations or law firm website videos featuring the voice of the executive. Why is deepfake audio more flexible? Well, with video, there needs to be a baseline video. Not so with audio deepfakes. Once there is a credible audio profile created, the attacker can use "text-to-speech" software and create any script desired for the phony voice to read.

A year ago, creating deepfake audio was not easy or cheap—it took time and money, which is a bar to many attackers. The most advanced systems took just 20 minutes of audio to create a voice profile, but in most cases systems needed hundreds of hours of audio to create a credible deepfake audio. All of that has changed. Today, competent audio deepfakes can be generated from just a few minutes of audio.

No longer do you have to spend thousands of dollars training a very convincing deepfake audio model. There are many examples of very credible synthetic audio. Just jump over to the Vocal Synthesis YouTube channel and you'll be totally impressed at the quality and accuracy of Jimmy Carter reading the Navy Seals Copypasta. The audio was created entirely by a computer using text-to-speech software trained on the speech patterns of Jimmy Carter.

You can increase your success rate by adding additional environmental sounds. Background noise of some kind, traffic noises, for instance – that help gloss over imperfections in the audio. Real attacks we've seen thus far cleverly used background noise to mask imperfections, for example, simulating someone calling from a spotty cellular phone connection or being in a busy area with a lot of traffic.

These attacks are no longer hypothetical. Last year, a CEO was scammed out of $243,000 by an audio deepfake purporting to be the chief executive of the firm's German parent company. The convincing part was that the CEO recognized the slight German accent in his boss' voice.

Security company Pindrop analyzes voice interactions in order to help prevent bank fraud and has come across some pretty clever criminals using background sounds to throw off banking reps. "There's a fraudster we called Chicken Man who always had roosters going in the background. And there is one lady who used a baby crying in the background to essentially convince the call center agents, that 'hey, I am going through a tough time' to get sympathy." We've got to admit—these people are crafty.

How Can We Defeat Audio Deepfakes?

Those who can authorize wire payments might want to take a look at the audio footprints they have in the public realm to assess their degree of possible risk? If the risk is considerable, they should tighten requirements before funds are sent.

Naturally, researchers are working out ways to review the audio of a call authorizing the release of funds to assess the probability that it is real or fake. We are not sure yet about the implementation of a certification system for inter-organizational calls, but it's a possibility. Another possibility would use blockchain technology in combination with voice-over-IP (VoiP) calls in order provide caller authentication. Upper-level law firm personnel who have the authority to issue payments should review their available body of public audio to determine their level of risk and perhaps implement added verification requirements. Of course, the possibility that an attacker might engage a target in a phone or in-person conversation to obtain the voice data they need should also be considered as this takes its place among the more common AI cyberattacks. Yikes.

Researchers from Symantec have undertaken a new method of analyzing a call's audio – in the end producing a probability rating of its authenticity. Is it possible to employ a certification system for inter-organizational calls? Researchers are looking at that too. Can we use the technology of blockchain as a means of authentication? That is also being researched. As you can imagine, it is (as they say) complicated.

Defenses against these attacks summon all the rules of cybersecurity, including the wisdom of educating employees. There is little knowledge among employees about these deepfakes, particularly the audio fakes. After training, employees are much more apt to be suspicious of an unusual payment request or – another frequent ploy – a request for network access. While we wait for new technology, protection against these new AI cyberattacks ties in with basic cybersecurity in handling all forms of BEC and invoicing fraud—the foundation is employee education. Most employees are not currently aware of what deepfake audios are, let alone the possibility that faked audio can be used to simulate a call from a superior. Education can motivate an employee to question an unusual payment or network access request. Putting additional verification methods in place is wise.

There is both good and bad news when it comes to identifying fake audio. As previously mentioned, technology is making the voice fabrications better and better every day. High-fidelity audio can sound pretty convincing to the human ear. However, the longer the clip, the more likely you are to be suspicious when you hear slight imperfections. Be suspicious if the audio sounds a little too "clean" as if generated with professional equipment in a recording studio.

When Pindrop analyzed audio, the systems look for characteristics of human speech to determine if it is even physically possible to make the sounds heard. "When we look at synthesized audio, we sometimes see things and say, 'this could never have been generated by a human because the only person who could have generated this needs to have a seven-foot-long neck."

Baseline BEC defenses such as filtering and authentication frameworks for email can stop these attacks in their tracks by snagging phishing emails before being delivered. As always, require multi-factor wherever you can. We always advise law firms to require that an employee who receives a wire funds request call the authorizing party back – at a known good number – never at a number given in the audio message. Verify everything!

New Targets

Original deepfakes targeted politicians, public figures and celebrities since they took hours of video showing the target's face to generate believable output. That's all changed. Today's advanced artificial

intelligence and machine learning can generate believable deepfakes using a single picture of the target and only five seconds of their voice. This means that anyone can be a victim if you use any social media or participate in video conference calls.

Deepfake ransomware is the latest trend. Security company, Trend Micro, reported that deepfakes can be used to blackmail people into paying significant ransom fees or turning over sensitive company information. Think of it as the next generation sextortion scam. Cybercriminals are monetizing the deepfake scam. Trend Micro stated, "A real image or video would be unnecessary. Virtually blackmailing individuals is more efficient because cybercriminals wouldn't need to socially engineer someone into a compromising position. The attacker starts with an incriminating Deepfake video, created from videos of the victim's face and samples of their voice collected from social media accounts. To further pressure the victim, the attacker could start a countdown timer and include a link to a fake video. If the victim does not pay before the deadline, all contacts in their address books will receive the link." Scary stuff.

Evidentiary Impact

Fabrication of email and text messages has been around for several decades. It is fairly easy to spoof an email or make a text message appear to come from someone other than the real person. The quick and simple way is to change the contact name of the phone number in your phone and have your accomplice (the one that really has the phone number) send you text messages supporting your fraud. That's exactly what happened in a recent California case.

The landscape has now evolved to video manipulation. Pretty much everyone has a smartphone with a camera capable of high-resolution recording. Even home security cameras can deliver 4K resolution and are not very expensive. Armed with video evidence, the footage can be modified to tell a different story. In other words, think of it as a type of video deepfake. The tools are plentiful and commercially available. Some are even free. As Jude Egan, a certified Family law Specialist certified by the California State Bar Board of Legal Specialization, stated, "As it gets simpler to manipulate video footage, I will have no real way of knowing what footage has been manipulated and what has not. This is especially true for footage taken when my client – or anyone I can find as an actual witness – was not present when the footage was taken, such as the other party getting into a public brawl or being intoxicated or under the influence of drugs."

Electronic evidence has always been a thorny subject with jury trials. Many jurors expect there to be scientific testing and testimony (e.g. DNA) to support the evidence. It is known as the CSI effect. Jurors want to hear from an expert just like on the CSI TV series. The situation isn't very different with the potential for deepfakes. Jurors or attorneys may think they are dealing with a deepfake when the video is totally authentic. The suspicions may now require each party to have an expert testify that the evidence is real or fabricated. We believe this is the next generation of the CSI effect. If one party doesn't put an expert on the stand, jurors may not trust the evidence. In other words, "Why didn't you have an expert testify that there was no manipulation of the video? That must mean the video is a deepfake." Unfortunately, our beliefs are supported by Riana Pfefferkorn, associate director of surveillance and cybersecurity at Stanford's Center for Internet and Society. She said, "This is dangerous in the courtroom context because the ultimate goal is to seek out truth. My fear is that the cultural worry could be weaponized to discredit [videos] and lead jurors to discount evidence that is authentic." The burden would then shift to the party that introduced the evidence to prove that it is authentic. This is a similar phenomenon that occurs with e-discovery, where costs can skyrocket when one party makes unreasonable requests dealing with electronic evidence.

Even Hany Farid opined, “I expect that in this and other realms, the rise of AI-synthesized content will increase the likelihood and efficacy of those claiming that real content is fake.”

Final Thoughts

We tried to keep the information above as readable and informative as possible. The truth is, very few lawyers understand the risks of deepfake audios and videos and how to address them. More and more, we are asking law firms to accept a homework assignment, which is now an important piece of the ever-evolving cybersecurity threats. While it is fairly inexpensive to generate credible deepfake audio today, generating credible video deepfakes is a lot more difficult. However, as the cost of computational horsepower goes down, we’re sure that video deepfakes will become more affordable to the masses.

Sharon D. Nelson is a practicing attorney and the president of Sensei Enterprises, Inc. She is a past president of the Virginia State Bar, the Fairfax Bar Association and the Fairfax Law Foundation. She a co-author of 18 books published by the ABA. snelson@senseient.com

John W. Simek is vice president of Sensei Enterprises, Inc. He is a Certified Information Systems Security Professional, Certified Ethical Hacker, and a nationally known expert in the area of digital forensics. He and Sharon provide legal technology, cybersecurity and digital forensics services from their Fairfax, Virginia firm. jsimek@senseient.com.