Chapter 4. Authenticity: Unique Challenges

This chapter is current through December 2021.

4.A. Deepfakes

In recent years, we have seen the proliferation of a new kind of manipulated media called “deepfakes.” In simple terms, deepfakes are convincing media forgeries created by computers. In these fake videos, audio recordings, and images, people often appear to be doing something they have never done or saying things they have never said.

4.A.1. Deepfakes Defined

The word “deepfake” is a portmanteau of two words, “deep learning” and “fake.”[1] Deep learning is a branch of artificial intelligence (AI) that mimics the workings of the human brain in processing data.[2]

Deepfakes are synthetic media (text, images, audio, or video) that are either manipulated or wholly generated by AI.[3] Deepfakes are made using computer systems called “generative adversarial networks” (GANs). These networks rely on an available data set of images to generate an accurate image or video. In a GAN, one algorithm (the generator) creates content modeled on source data, while a second algorithm (the discriminator) tries to spot the artificial content.[4] The competition between the two networks produces a better-and-better fake until the discriminator can no longer identify the forgery.

Recent examples of deepfakes include a 2018 video of President Barack Obama talking about deepfakes. In fact, the producers took an original video of President Obama and used AI to alter his lip movements and to overlay the voice of actor Jordan Peele.[5] And in 2020, MIT researchers produced a video in which President Richard Nixon seemed to announce that the Apollo 11 moon landing had ended in disaster. There, film from a different Nixon speech was edited with deepfake technology to alter Nixon's voice and facial movements to make it appear as though he gave a speech he never did.[6]

4.A.2. Dangers Of Deepfakes

The technology used to make deepfakes is getting better and more accessible. Already, there are apps that virtually anyone can download on their phone to place a celebrity's face into the user's video selfies.[7] The number of deepfakes is increasingly rapidly. According to an analysis by the AI company Sentinel, as of 2020, there were more than 100 million deepfake videos online, which represented a year-over-year growth rate of 6,820%.[8] . A separate study using different methodology estimated in July 2019 that the number of deepfake videos on the Internet was doubling every six months.[9] Soon, the ability to create ever more convincing fake media showing people doing and saying things that did not occur in reality will be widespread.

Deepfakes targeting women are already prevalent. According to a 2019 study, over 95% of all deepfake videos available on the internet are of nonconsensual pornography—that is, a nonconsenting woman's face (whether a celebrity or an average person) is placed on the body of a pornographic performer.[10] The technology that makes those kinds of videos can also be used to create convincing deepfakes of politicians, public officials, and business people, with potentially serious repercussions for society.

The propagation of deepfakes could also foster what law Professors Bobby Chesney and Danielle Citron have called the “Liar's Dividend,”[11] wherein individuals successfully deny the authenticity of genuine media by claiming that the content is a deepfake. The potency of the liar's dividend will grow as more people learn about the quality and availability of this technology.

4.A.3. Deepfakes In The Courtroom

Highly realistic media manipulation poses a new challenge to the authenticity of evidence admitted into court.[12] As the forgeries become more realistic and their presence more commonplace, it will become harder to detect manipulated multimedia evidence; judges will have to rule on how such evidence can be authenticated, and jurors may become suspicious of media in general, according to experts.[13]

The Federal Rules of Evidence allow for the authentication of records "generated by an electronic process or system that produces an accurate result," if "shown by a certification of a qualified person" in a manner set forth by the rules.[14] The rules require advance notice to the other side and the opportunity to challenge the records.[15] If the certification requirements are met, a party need not call testifying witness at trial to establish authenticity.

Practitioners should be alert for opportunities to challenge admissibility on the basis that the evidence may have been faked. What once may have been accepted without question, could now be vulnerable to attack. Similarly, judges should prepare for competing expert testimony over the authenticity of videos and audio recordings—media that would previously have been considered unimpeachable.[16]

For example, in a recent case, an attorney in the United Arab Emirates claimed that a manipulated audio recording was used in a child custody case brought in the United Kingdom to discredit the child's UAE-based father. The mother in the case relied on software and online tutorials to manipulate an audio recording of the father to include words he did not say to make it sound as though he was threatening the mother. According to reports, forensic experts were able to show the court how the media had been manipulated.[17]

Riana Pfefferkorn, Research Scholar at the Stanford Internet Observatory, suggests attorneys budget for digital forensic experts and witnesses and learn to recognize the telltale signs of manipulated media.[18] "Deepfakes will soon make trial attorneys' and judges' jobs more difficult," she writes, "[l]awyers will have to exercise greater diligence in verifying the authenticity of video evidence."[19]

In an August 2020 article, California family law specialist M. Jude Egan pointed out that California judges consider petitions for Domestic Violence Temporary Restraining Orders (DVTROs) without notice to the other party and often rely on digital evidence submitted without third-party verification. The rise of manipulated media and the potency of DVTROs "should give lawyers and judges pause when reviewing this type of information in a DVTRO request and particularly with video evidence where detection of manipulation is getting more difficult at the same time as it gets cheaper and easier," Egan wrote. He suggested judges not extend DVTROs without proof that such evidence is legitimate, that parties relying on electronic evidence provide as much corroborating evidence as possible, and that parties found to be faking such evidence be prosecuted for perjury.[20]

Issues concerning the authenticity or inauthenticity of manipulated media, like deepfakes, are beginning to arise in other litigation contexts as well.

For example, in *People v. Beckley* the California Court of Appeal held that the prosecution's failure to authenticate a photograph downloaded from the internet should have barred its admission.[21] Although the court considered the admission of the photograph harmless, it held that under California Rules of Evidence 250 and 1401, a photograph is a "writing" and "[a]uthentication of a writing is required before it may be received in evidence."[22] The court observed that it is well settled "that the testimony of a person who was present at the time a film was made that it accurately depicts what it purports to show is a legally sufficient foundation for its admission into evidence" and that authentication "may be provided by the aid of expert testimony."[23]

In *Beckley*, the prosecution alleged that the photo showed a defense witness flashing a gang sign, undermining her credibility. But the investigating detective could not testify from his personal knowledge that the photograph truthfully portrayed the witness flashing the gang sign, and no expert testified that the picture was not a "composite or faked"

photograph.[24] The court observed that "[s]uch expert testimony" is "critical today to prevent the admission of manipulated images," especially when "[r]ecent experience shows that digital photographs can be changed to produce false images. Indeed, with the advent of computer software programs such as Adobe Photoshop it does not always take skill, experience, or even cognizance to alter a digital photo."[25]

A recent case in Colorado likewise illustrates the growing skepticism some courts harbor toward media that once was thought to be nearly unassailable. In *People v. Gonzales*, the Colorado Court of Appeals held that under Colorado Rule of Evidence 901—which requires the trial court to consider all the circumstances surrounding proffered evidence—the authentication requirement for admission into trial of a voicemail allegedly left by the defendant for the victim was satisfied.[26] In that case, a prosecution for murder, the recording had been found in the victim's house by the victim's sister. The police officer who interrogated the defendant testified that the defendant's voice was heard on the voicemail, and the defendant did not claim the recording was falsified or manipulated.[27] The court held that these facts satisfied Rule 901's "flexible, factual inquiry."[28] In an earlier age, the matter may have been left there. But, instead, the Supreme Court of Colorado in October 2019 granted certiorari to consider whether the admissibility bar for voice recordings was too low, and "[w]hether a voice recording may be admitted into evidence when there is no witness who can vouch for either the accuracy of the recording's contents or the reliability of the recording process."[29]

In September 2020, the Colorado Supreme Court affirmed the defendant's conviction, holding a voice recording may be admitted into evidence without such witnesses. The court held that "[i]n the absence of evidence suggesting that a proffered voice recording has been altered or fabricated…a proponent may authenticate a recording by presenting evidence sufficient to support a finding that it is what the proponent claims. Once this prima facie burden is met, authenticity becomes a question for the factfinder…."[30]

In *People v. Foreman*, an Illinois defendant asked a state appeals court to be particularly suspicious of recorded communications because of the availability of deepfake technology. In that case, the defendant's murder conviction rested in part on recorded prison telephone calls. The defendant claimed that the trial court erred in ruling that the State had laid a sufficient foundation for the reliability and accuracy of the recorded jail communications.[31] The defendant argued that "in the age of so-called deep-fake videos and easily-manipulated audio recordings, improperly-authenticated recorded communications should be inherently suspect."[32]

In July 2020, the Appellate Court of Illinois rejected the challenge to the recordings' authenticity and affirmed the conviction and the sentence. The court held that under Illinois' silent-witness theory the State had laid a proper foundation for the admission of the audio recordings because the prison's administration operations captain testified as to the operation of the prison phone system.[33] The court also discarded the argument that "recent technological advancements render all recordings suspect, because they can be easily manipulated." It held that, "[i]n the absence of any evidence of tampering or other such manipulation in this case, there are no foundational issues with the recordings."[34]

And in June 2020, in *Newell v. United States*, an appellant invoked the specter of deepfake technology to challenge his conviction on federal child pornography charges. The *pro se* appellant challenged the legal definition of child pornography, defined as visual depictions involving "the use of a minor engaging in sexually explicit conduct,"[35] because "in the era of deep fakes…jurors can no longer differentiate between real versus fake human faces."[36] Citing research that some individuals cannot distinguish between authentic and manufactured deepfake images of human faces, the appellant argued that "it is no longer acceptable for a prosecutor's presentation of images alone to constitute sufficient evidence for a jury to determine that an actual child is depicted in pornography."[37] The appellant contended that, "[i]n the deep fake era, the only time that a defendant could knowingly possess images of an actual child would be [when] the minor was identifiable to the defendant."[38]

The Government challenged that assertion, arguing that under Circuit precedent, "a prosecutor's presentation of images alone constitutes sufficient evidence for a jury to determine that an actual child is depicted in pornography."[39] The Court of Appeals affirmed the lower-court's judgment dismissing the appellant's appeal without reaching his arguments on deepfake

technology.[40]

4.A.4. Detecting Deepfakes

Technologists are working on reliable ways to detect deepfakes. These technologies fall into roughly two categories.[41] The first method attempts to detect the fake media after it is created. For example, that technology was relied upon recently to reveal that the headshots used by supposed op-ed authors were fake and created by AI.[42] The second method verifies photographs at the "point of capture" in such a way that they cannot be altered or modified after the fact.[43] Efforts to improve and distribute these countermeasure technologies, like efforts to improve and distribute deepfake technology itself, are ongoing.

The American College of Trial Lawyers Handbook of Electronic Evidence

---

[1] Jeffery DelViscio, *A Nixon Deepfake, a 'Moon Disaster' Speech and an Information Ecosystem at Risk*, Scientific American, July 20, 2020, https://www.scientificamerican.com/article/a-nixon-deepfake-a-moon-disaster-speech-and-an-information-ecosystem-at-risk1/.

[2] "Deep Learning", Investopedia, https://www.investopedia.com/terms/d/deep-learning.asp.

[3] *See generally* Nina Schick, Deepfakes: The Coming Infocalypse 8 (2020).

[4] Sarah Basford, *What Deepfakes Actually Are*, Gizmodo (Jul. 31, 2020), https://www.gizmodo.com.au/2020/07/what-are-deepfakes/; Meredith Somers, *Deepfakes, Explained*, MIT (Jul. 21, 2020), https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained

[5] Basford, *What Deepfakes Actually Are*.

[6] DelViscio, *A Nixon Deepfake*.

[7] Matthew F. Ferraro, Jason C. Chipman & Stephen W. Preston, *Identifying the Legal and Business Risks of Disinformation and Deepfakes: What Every Business Needs to Know*, 6 Pratt's Privacy and Cybersecurity Law Report 142, 152 (2020) (describing "Impressions.app").

[8] Sentinel, Deepfakes 2020: The Tipping Point 7 (2020), https://thesentinel.ai/media/Deepfakes%202020:%20The%20Tipping%20Point,%20Sentinel.pdf .

[9] Henry Ajder, *The State of Deepfakes 2019: Landscape, Threats, and Impact*, Sensity AI (Sept. 2019), https://sensity.ai/reports/#.

[10] Giorgio Patrini, *Mapping the Deepfake Landscape*, Sensity (Oct. 7, 2019), https://sensity.ai/mapping-the-deepfake-landscape/.

[11] Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107

California Law Review 1753, 1785-1786 (2019).

[12] *See* Ferraro *et al.*, *Identifying the Legal and Business Risks*, at 151.

[13] Matt Reynolds, *Courts and Lawyers Struggle with Growing Prevalence of Deepfakes*, ABA Journal (June 9, 2020), https://www.abajournal.com/web/article/courts-and-lawyers-struggle-with-growing-prevalence-of-deepfakes (quoting Riana Pfefferkorn).

[14] Federal Rule of Evidence 902(13); *see also* Fed. R. Evid. 902(14).

[15] Federal Rule of Evidence 902(11) and 902(12)

[16] Ferraro *et al.*, *Identifying the Legal and Business Risks*, at 151; Theodore F. Claypoole, *AI and Evidence: Let's Start to Worry*, The National Law Review (Nov. 14, 2019), https://www.natlawreview.com/article/ai-and-evidence-let-s-start-to-worry; Riana Pfefferkorn, *Too Good to Be True?*, NW Lawyer (September 2019), at 22, 24; Reynolds, *Courts and Lawyers Struggle*.

[17] Ferraro *et al.*, *Identifying the Legal and Business Risks*, at 151-152 (citing reports).

[18] Reynolds, *Courts and Lawyers Struggle* (citing Pfefferkorn).

[19] Riana Pfefferkorn, *"Deepfakes" in the Courtroom*, 29 Pub. Int. L. J. 245, 275 (2020).

[20] M. Jude Egan, *Deep Fakes in Divorce Court: Manipulated Electronic Evidence and What to Do About It*, The Recorder (Aug. 20, 2020), https://www.law.com/therecorder/2020/08/20/deep-fakes-in-divorce-court-manipulated-electronic-evidence-and-what-to-do-about-it/.

[21] *People v. Beckley*, 110 Cal. Rptr.3d 362 (Cal. Ct. App. 2010).

[22] *Id.* at 366 (internal citations and quotation marks omitted).

[23] *Id.* (internal citations and quotation marks omitted).

[24] *Id.* (internal citations and quotation marks omitted).

[25] *Id.* at 366-367 (internal citations and quotation marks omitted).

[26] *People v. Gonzales*, 2019 WL 1087008, at *6 (Colo. App., 2019).

[27] *Id.*

[28] *Id.* at *4 (Colo. App., 2019).

[29] *Gonzales v. People*, No. 19SC292, *2019 WL 5196949*, at *1 (Colo., Oct. 15, 2019).

[30] *Gonzales v. People*, 471 P.3d 1059, 1062 (Colo., 2020) (internal citation omitted).

[31] *People v. Foreman*, *2020 WL 4037351*, at *13-14 (Ill. App. Ct.., 2020).

[32] *Id.* at *15.

[33] *Id.* at *14, 16.

[34] *Id.* at *17.

[35] 18 U.S.C. § 2256(8)(A).

[36] Opening Brief, *Newell v. United States*, No. 19-56522, *2020 WL 554527*, at *5 (9th Cir., Jan. 28, 2020).

[37] *Id.* at *6.

[38] *Id.* at *8.

[39] Appellees' Answering Brief, *Newell v. United States*, No. 19-56522, *2020 WL 3493673*, at *37 (9th Cir., June 18, 2020).

[40] *Newell v. Garland*, 846 F. App'x. 523, 524 (9th Cir., 2021), *cert. denied*, *2021 WL 4507888* (U.S., Oct. 4, 2021).

[41] *See generally* Matthew F. Ferraro, *Decoding Deepfakes*, National Security Institute Backgrounder (2020), https://nationalsecurity.gmu.edu/ddf/.

[42] Raphael Satter, *Deepfake Used to Attack Activist Couple Shows New Disinformation Frontier*, Reuters (Jul. 15, 2020), https://www.reuters.com/article/us-cyber-deepfake-activist/deepfake-used-to-attack-activist-couple-shows-new-disinformation-frontier-idUSKCN24G15E.

[43] Mounir Ibrahim, *To Beat Deepfakes, We Need to Prove What is Real. Here's How*, World Economic Forum (Mar. 23, 2020), https://www.weforum.org/agenda/2020/03/how-to-make-better-decisions-in-the-deepfake-era/.