DEALING WITH DEEPFAKES

BY JUDGE BERNICE B. DONALD
AND RONALD J. HEDGES, J.D.

Advances in technology have created several new software tools that allow for the creation of fabricated content. These “deepfakes” may represent a social media phenomenon today, however, they will inevitably go beyond that status and become a feature in litigation, both civil and criminal.

The fact that deepfake content is falsified and artificially created means that victims of deepfakes will have difficulty claiming that there was a privacy violation. This article will consider how discovery of deepfakes may be conducted in litigation and then address “practical” issues such as admissibility of evidence about deepfakes, the role of experts, and costs.

So, what is a deepfake? The term “deepfake” is a combination of the words “deep learning” (a branch of artificial technology) and “fake,” and is attributed to Reddit user who shared videos of his face-swapping hobby. Hubert Davis, How Deepfake Technology Actually Works, SCREENRANT, Feb. 21, 2020, https://screenrant.com/deepfake-videos-explained-how/. These are pictures, videos, and audio recordings that look and sound real, but have actually been manipulated using artificial intelligence (“AI”) in order to make people appear to do and say things that they never did or said. The underlying technology can replace faces, manipulate facial expressions, synthesize faces, and synthesize speech.” See U.S. GOV’T ACCOUNTABILITY OFFICE, SCI. & TECH. SPOTLIGHT: DEEPFAKES (Feb. 2020), https://www.gao.gov/assets/710/704774.pdf.

Common types of deepfakes include (1) face swapping (the face of one person is placed on to the body of a different person; (2) Lip-sync (an algorithm takes a video of a person speaking and alters the person's lip movements to match the lip movements of a different audio track); and (3) Puppet-master (an image of a person's body is swapped with an image of another person's body, enabling the coder to "hijack people's entire bodies and animate them" as he wishes). The Spooky Underpinnings of Deepfakes, DIGITAL MOUNTAIN, (Sept. 9, 2020, 1:52 PM), http://digitalmountain.com/enews/FALL_2019_Article1.pdf.

1 The terms “deepfake” and “deep fake” has been used interchangeably in academia and media. For consistency purposes, we refer to the term as a single word.
Deepfakes are created by feeding large data sets of images into the artificial neural networks, which train these computer systems to recognize and reconstruct facial biometrics and other patterns. Two of the AI technologies used to create deepfakes are (1) Autoencoders (artificial neural networks that are trained to reconstruct input from a simpler representation), and (2) Generative Adversarial Networks (GANs), made up of two competing artificial neural networks, one trying to produce a fake, the other trying to detect it. This training continues over many cycles, resulting in a more plausible rendering of, for example, faces in video. GANs generally produce more convincing deepfakes but are more difficult to use. See U.S. GOV'T ACCOUNTABILITY OFFICE, SCI. & TECH. SPOTLIGHT: DEEPFAKES (Feb. 2020), https://www.gao.gov/assets/710/704774.pdf.

PRIVACY AND CYBERSECURITY CHALLENGES POSED BY DEEPFAKES

Social media has fueled the spread of deepfake videos and images. While it had been difficult to create deepfakes without advanced training in computer science, new software has enabled the “guy in his mother’s basement with a PC” to be able to create them. See Kristen Dold, Face-Swapping Porn: How a Creepy Internet Trend Could Threaten Democracy, ROLLING STONE (Aug. 31, 2020, 3:52 PM), https://www.rollingstone.com/culture/culture-features/face-swapping-porn-how-a-creepy-internet-trend-could-threaten-democracy-629275. Although it is true that social media has begun screening content for possible deepfakes, it only catches about “two-thirds” of them. Will Knight, Deepfakes Aren’t Very Good. Nor are the Tools to Detect Them, WIRED, (June 12, 2020), https://www.wired.com/story/deepfakes-not-very-good-nor-tools-detect/.

Beyond the social media context, business and governmental entities create, collect, and store ever-increasing volumes of personal information (think of images for driver’s licenses as an example) that may become subjects of data breaches. This availability — along with financial or other incentives for capture and “repurposing” of personal images — raises significant privacy-related issues that State laws, such as the California Consumer Privacy Act, the Illinois Biometric Information Protection Act, and the New York SHIELD Act are intended to address. Cal. Civ. Code, § 1798.100 et seq., (2019); 40 Ill. Comp. Stat. Ann. §14/15 (2018); N.Y. Gen Bus. Law§ 899-bb (2019). States have also begun responding to civil and criminal challenges caused by deepfakes, but so far, legislative change has been slow.

DEEPFAKES IN LITIGATION

Assuming that a deepfake may give rise to liability under a tort, breach of contract or privacy law theory, and the creator of the deepfake can be identified, several other issues remain.

First, triggering the duty to preserve will be essential. The lawyer representing the victim of a deepfake will need to send a notice to the defendant, informing her of the client’s intent to commence litigation against that defendant. This notice may be sufficient to trigger a duty to preserve any relevant deepfake image within the defendant’s possession, custody, or control. Less clear, in the context of a deepfake, are the categories of data and volume of data that the defendant has a duty to preserve. Questions include:

- What does it mean to preserve the image?
- Must the defendant preserve any unaltered images that were used to create the altered and objectionable one?
- Must the defendant preserve any metadata associated with those images?
- Must the defendant preserve all the data that was used to train the AI that resulted in the deepfake?
- Must the defendant preserve the AI tools that were used to create the altered images?

As for the plaintiff’s duty to preserve, if the plaintiff or her lawyer retained an expert who used an AI tool to identify a video as a deepfake, she may have a duty to preserve the
data and AI tools that formed the basis for her claim that the video at issue is a deepfake. The questions about the scope of this duty are similar to the questions about the scope of the defendant’s duty to preserve listed above. Given the scarcity of caselaw on deepfakes, the lawyers and judges involved in these cases will be breaking new ground.

Second, the need for protective orders may grow: In order to establish the elements of a claim under the California deepfake law, or based on tort, breach of contract, or privacy law theories, it may be necessary for the plaintiff to offer testimony about how the underlying technologies at the heart of the claim function. Protective orders may be necessary if the defendant contends that the technology behind the alleged deepfake is proprietary in nature and if the plaintiff takes the position that the technology used by her expert to identify the recording at issue as a deepfake is proprietary. The court will then need to decide what provisions and conditions should be imposed on the parties to ensure the confidentiality and security of the data describing the operations of these technologies.

Third, admissibility of evidence may become highly contested: If a civil action involving a deepfake proceeds to trial, a central question will presumably be the admissibility of the deepfake. The analysis of that issue will involve answering the following questions:

• Is the deepfake relevant?
• Has the deepfake been authenticated?
• Is the deepfake hearsay?
• Is the deepfake an “original”?
• Would admitting the deepfake into evidence be unduly prejudicial?

Authentication is often the most difficult hurdle to the admissibility of electronic evidence. For example, digital photography was rarely doubted for authenticity reasons. The presence of deepfakes may require parties to lay a more elaborate foundation, and for courts to impose a higher burden of admissibility on digital content. See 3 TRIAL HANDBOOK FOR ARK. LAWYERS § 62:4 (2019-2020 ed.). Some types of evidence, such as those listed in Federal Rule of Evidence 902, can be self-authenticated through the declaration of a qualified person. FED. R. EVID. 902. Two categories of evidence listed in Rule 902 that could apply to deepfakes are: (1) “record[s] generated by an electronic process or system that produces an accurate result,” FED. R. EVID. 902(13), and (2) “data copied from an electronic device... if authenticated by a process of digital identification.” FED. R. EVID. 902(14).

If the deepfake does not fall under one of the categories in Rule 902, the party seeking to admit the deepfake will need offer the testimony of either a layperson (under FED. R. EVID. 701) or an expert witness (under FED. R. EVID. 702) in order to authenticate the deepfake. Authenticating witnesses, such as expert witnesses, may not be able to identify that the images or videos are altered. See John P. LaMonaca, A Break from Reality: Modernizing Authentication Standards for Digital Video Evidence in the Era of Deepfakes, 69 AM. U. L. REV. 1945, 1977 (2020). The opposing party can challenge the authenticity of the deepfake, and is likely to need information about the technologies that created it as well as the assistance of a forensic expert. See Committee Note to FED. R. EVID. 902(13).

In determining the admissibility of the alleged deepfake, the judge acts as gatekeeper. If the judge determines that the deepfake can be admitted into evidence, it is then up to the jury to resolve factual questions, such as whether the image, video or audio recording at issue is a deepfake and whether a defendant should be held responsible for the creation and/or distribution of the deepfake under civil or criminal law.

DEEPFAKES AND DISCOVERY

When and how deepfakes may be the subject of discovery has not been defined either. Generally, parties to a litigation may obtain discovery regarding any nonprivileged matter that is relevant to a claim or
defense in the action, and proportional to the needs of the case. Factors considered in the proportionality determination include: (1) the importance of the issues at stake in the action; (2) the amount in controversy; (3) the parties’ relative access to relevant information; (4) the resources of the parties; (5) the importance of the discovery that is being sought in resolving the issues in the case; and (6) whether the burden or expense of the proposed discovery outweighs its likely benefit. FED. R. CIV. P. 26(b)(1). Thus, in a civil case involving a deepfake, discovery into the deepfake will be allowed, subject to the proportionality analysis discussed above.

Another determination that might have to be made during the discovery process is whether the data being sought is within the “possession, custody, or control” of the other party. FED. R. CIV. P. 34(a)(1). If it is not, the party seeking the information may have to subpoena it, and that subpoena will have to satisfy the relevance and proportionality requirements under FED. R. CIV. P. 26(b)(1).

The court may also be called on to decide how much discovery each party is entitled to conduct about the design and workings of the technology used by the other party. For example, if the plaintiff requests extensive information about the AI, GANs, and data sets used to create the alleged deepfake and notices the depositions of a significant number of defendant’s employees who designed or operated the technology that led to the deepfake, the court may need to conduct a proportionality analysis to determine the categories and volume of data and number of depositions to which the plaintiff is entitled at that stage of the case.

In summary, discovery of deepfakes may be neither simple nor inexpensive. Moreover, it will likely require an attorney competent within the meaning of Model Rule of Professional Conduct 1.1 and its State equivalents to undertake that discovery.

Acknowledgements
Special thanks to Naira Umarov, Judicial Law Clerk to the Hon. Bernice B. Donald, U.S. Court of Appeals for the Sixth Circuit for her assistance with this article.