Deepfake dilemma

What will it mean to be defamed in a world where every aspect of a person can be fabricated by an AI, asks **Jason Haas**

The spread of non-AI tools to manipulate recordings already poses serious questions of how individuals can protect their reputations. In recent incidents, Speaker of the House Nancy Pelosi's voice was altered to make her sound intoxicated,[1] while another video made it falsely appear that a reporter struck a White House intern.[2] A photograph of former NFL quarterback Colin Kaepernick in a political fund-raising email was changed to darken his skin.[3] While these amateurish attempts were quickly identified as frauds, the development of AI will eventually make the tools to create sophisticated fake recordings widely available, fakes that will become increasingly difficult to disprove. The recent craze for FaceApp, an AI-fueled application that allows you to easily change the age of a person in a picture, is just one example of what emerging technology will make possible.

Since 2017, "deepfake" has been used to refer to a machine-learning technique that transforms existing video recordings to create realistic but false depictions of events. This practice originated with the swapping of celebrities' faces into pornographic videos and "revenge porn", but deepfakes are now moving out of the realm of voyeurism and into other areas of society. Of course, there are legitimate commercial uses for such technology. Yet, when innocent parties are injured by deepfakes, they will turn to the law and such doctrines as defamation, "false light" invasion of privacy, and rights of publicity for recourse. This article addresses the impact of deepfakes only on the law of defamation.

Current law

While defamation in the US is regulated by the states, the tort typically requires an unprivileged publication of a false and defamatory statement concerning another person where harm to a reputation can be presumed or "special harm" can be shown. A defamatory "statement" can include a photograph or movie, and altered photographs can be defamatory even without accompanying text.[4]

When public officials or figures claim defamation, they must satisfy the Supreme Court's "actual malice" standard adopted in *New York Times Co v Sullivan*.[5] This requires proof that statements were intentionally false or made with a "reckless disregard for the truth", which can require showing a "high degree of awareness of ... probable falsity" or that the publisher "entertained serious doubts as to the truth of his publication".[6] Surviving summary judgment requires "clear and convincing" evidence of actual malice.[7]

Private figures must only show actual malice to recover presumed or punitive damages for defamation implicating a "matter of public concern".[8]

Identity

Merely identifying who created a deepfake may prove impossible. While most defamation cases have identifiable defendants, a deepfake can spread anonymously on the internet. The original "deepfake" was a Reddit user who created the first face-swapping porn videos and remains unknown to this day.[9] Even expensive discovery measures may prove inadequate to identify a deepfake creator, leaving a plaintiff's only possible recourse to sue republishers.

Falsity

When a deepfake can be compared to an original recording, proving it a fake should be straightforward. However, as AI technology makes it possible to create realistic images and video from "whole cloth", this task will become increasingly difficult.

Computer scientists have developed tools to detect deepfakes by examining details of light and shadows, movement, facial features and clothing.[10] But as the quality of deepfakes improves, current detection measures will become less effective. New tools will be needed. Proving a recording false may often require complex (and expensive) expert testimony.

Actual malice

Courts frequently invoke the actual malice standard to dispense of defamation claims before trial. Proving knowledge of falsity (or reckless disregard of truth) imposes a high burden, particularly against media defendants reporting the allegedly-defamatory statements of others.

The actual malice standard will rarely protect deepfake creators, as evidence that a recording is a deepfake also proves their knowledge of falsity. While some creators might be able to argue their work constitute protected parodies, and others that their edits do not materially change

the original recordings' meaning or impact,[11] these arguments appear unlikely to aid many.

For media defendants, the real impact of deepfakes will be on their duty to investigate. Negligently publishing false information about private individuals is sufficient to impose liability on media defendants.[12] Meanwhile, republishing a deepfake relating to a public figure could show actual malice if there were obvious reasons to doubt its accuracy prior to publication.[13] As deepfakes becomes more widespread, media will be expected to apply greater scrutiny before publishing videos of questionable provenance or legitimacy. If a simple pre-publication review could have detected tampering, courts could easily find the lack of one to be an "extreme departure from the standards of investigation and reporting ordinarily adhered to by responsible publishers."[14] In time, running all footage through deepfake-detection programs could (and should) become an industry standard, and a failure to use such tools could significantly increase the risk of liability. Once there is a credible assertion that a recording is a deepfake, any further publication certainly risks a finding of actual malice.

"Courts may not wish to send every deepfake case to a jury. Some materiality or other standard will be needed to weed out less meritorious claims."

Defamatory meaning

Perhaps the most novel implications of deepfakes rests in the question of whether particular content is defamatory. Although states use many different formulations, the essence of defamation is that a publication must be so "significantly injurious to reputation to actually cause persons to think less of the plaintiff…"[15] Making this determination can already be a complex process with videos, as courts recognise that "a clever amalgamation of half-truths and opinion-like statements, adorned with orchestrated images and dramatic audio accompaniment, can be devastating when packaged in the powerful television medium."[16]

Deepfakes add a new level of complexity. While obvious cases like the Pelosi video should be no challenge, deepfakes will enable more subtle changes to be made. For example, is it defamatory to depict a person with darker skin? Does it harm a person's reputation to be presented as older than he actually is? Does altering a person's voice defame her?

To answer these questions, some courts might find a deepfake is "reasonably susceptible" to a defamatory meaning simply because a recording was intentionally altered. Others may consider alleged damages to help assess defamatory meaning, particularly when "special harm" must be shown. But deepfakes will require courts to confront questions about the defamatory meaning of changes to aspects of a person – such as age, facial features, and voice – for which little precedent may exist. Courts may not wish to send every deepfake case to a jury. Some materiality or other standard will be needed to weed out less meritorious claims.

Courts may also prove unwilling to find defamation based on changes to legally-protected characteristics, such as skin colour, age or nationality. If the government cannot discriminate based on such differences, and the law protects people with those characteristics, then how can courts find a plaintiff has been defamed by being falsely depicted in that manner? "Courts will not condone theories of recovery which promote or effectuate discriminatory conduct."[17]

Section 230

Online platforms receive substantial protection from section 230 of the Communications Decency Act, and this should extend to postings of deepfakes.[18] However, challenges to this protection are growing. The practice of some platforms to allow unquestionably defamatory videos, such as the Pelosi video, to remain on their websites,[19] will increase the pressure on Congress to alter section 230.

Footnotes

1. https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/?utm_term=.63d0de546d63
2. https://time.com/5449401/jim-acosta-cnn-trump-video/
3. https://www.miamiherald.com/news/politics-government/national-politics/article232550702.html
4. See eg, Cal Civil Code §45; *Muzikowski v Paramount Pictures Corp,* 322 F.3d 918, 924-927 (7th Cir 2003) (movie under Illinois law); *Crump v Beckley Newspapers, Inc,* 173 W Va 699 (1983) ("well established that… libel… includes defamation through the publication of pictures or photographs"); *Kiesau v Bantz,* 686 NW.2d 164 (Iowa 2004)(doctored photo can be defamatory), overruled on other grounds by *Alcala v Marriott Int'l, Inc Eyeglasses,* 880 NW.2d 699 (Iowa 2016).
5. 376 US 254 (1964).
6. *Harte-Hanks Communications, Inc v Connaughton*, 491 US 657, 667 (1989).
7. *Anderson v Liberty Lobby, Inc,* 477 US 242, 257 (1986).
8. *Gertz v Robert Welch, Inc*, 418 US 323, 347-350 (1974); *Dun & Bradstreet, Inc v Greenmoss Builders, Inc,* 472 US 749, 763 (1985).
9. https://qz.com/1199850/google-gave-the-world-powerful-open-source-ai-tools-and-the-world-made-porn-with-them/.
10. https://www.washingtonpost.com/technology/2019/06/12/top-ai-researchers-race-detect-deepfake-videos-we-are-outgunned/?utm_term=.1b217a15fe35
11. See *Masson v New Yorker Magazine, Inc,* 501 US 496, 517 (1991).
12. *Mandel v Boston Phoenix, Inc,* 456 F.3d 198, 209 (2006).
13. See *St Amant v Thompson*, 390 US 727, 732 (1968).
14. See *Curtis Pub Co v Butts,* 388 US 130, 155 (1967).
15. 1 Law of Defamation § 4:1 (2d ed.).
16. *Corporate Training Unlimited, Inc v NBC, Inc,* 868 F.Supp 501, 507 (EDNY 1994).
17. *Polygram Records, Inc v Superior Court,* 170 Cal App.3d 543, 557 (1985); see also *Albright v Morton*, 321 F.Supp.2d 130, 136-139 (D Mass 2004) (holding allegation of homosexuality could not be defamatory and observing that misidentifying a white person as an African-American could not be defamation per se, "even if segments of the community still held profoundly racist attitudes").
18. See *Bennett v Google LLC,* 882 F.3d 1163, 1166 (DC Cir 2018) (a "provider… of an interactive computer service" is immune for online postings containing "information provided by another information content provider" when a complaint seeks to hold it liable as the publisher of the information).
19. https://www.cnbc.com/2019/05/24/fake-nancy-pelosi-video-remains-on-facebook-and-twitter.html

Author

Jason Haas is a litigator at Ervin Cohen & Jessup in Beverly Hills, California. He has over 18 years of experience in resolving complex commercial, organisational and IP disputes in the tech field and other industries.