The Place for Illusions: Deepfake Technology and the Challenges of Regulating Unreality

Lindsey Joost

Follow this and additional works at: https://scholarship.law.ufl.edu/jlpp

Recommended Citation

Available at: https://scholarship.law.ufl.edu/jlpp/vol33/iss2/7

This Article is brought to you for free and open access by UF Law Scholarship Repository. It has been accepted for inclusion in University of Florida Journal of Law & Public Policy by an authorized editor of UF Law Scholarship Repository. For more information, please contact rachel@law.ufl.edu.
THE PLACE FOR ILLUSIONS: DEEPFAKE TECHNOLOGY AND THE CHALLENGES OF REGULATING UNREALITY

Lindsey Joost*

Abstract

Existing laws are insufficient to address the harms caused by deepfakes. This Note explores the characteristics of deepfakes that make preventing both the misuse of the technology and its proliferation on social media and the Internet difficult. This Note will then continue by explaining the special potential harms that deepfakes pose. The third part of this Note will address how the existing legal framework fails, including how Section 230 of the Communications Decency Act serves as a barrier to redress. Finally, this Note suggests potential remedies for deepfake transgressions to reduce the potential harm that deepfakes can inflict.

INTRODUCTION ................................................................. 310

I. BACKGROUND ................................................................. 313
   A. What Are Deepfakes? ................................................... 313
   B. How Are They Used? .................................................. 316
   C. What Harms Do They Pose? ........................................ 317
      1. Threat to Individual Privacy .................................... 317
      2. Threat to National Security and Politics .................... 318

II. WHY CURRENT LEGAL FRAMEWORKS FAIL .................. 320
   A. Tort Law ................................................................. 321
   B. Copyright ............................................................. 323
   C. Nonconsensual Pornography Laws ............................ 325
   D. Section 230 of the Communications Decency Act .......... 328
   E. DEEP FAKES Accountability Act ............................... 330

III. RECOMMENDATIONS ..................................................... 331

CONCLUSION ...................................................................... 331

And one eye-witness weighs
More than ten hear-fays. Seeing is believing,
All the world o’er.
– Plautus

* J.D., University of Florida Levin College of Law, Class of 2023.

1. The commonplace saying is a rough translation from “pluris est oculatus testis unus quam auriti decem,” meaning “one witness with good eyes is worth more than ten witnesses with good ears.” See REGINE MAY, APELIUS AND DRAMA: THE ASS ON STAGE 55 (2006).
“Jack’s in charge of the choir. They can be—what do you want them to be?”
“Hunters.”
—Lord of the Flies

INTRODUCTION

After recent controversy, rapper Eminem has declared himself a feminist. Following the release of a poorly-received diss-track aimed at Facebook founder Mark Zuckerberg, the rapper released a new record—a “diss against the patriarchy”—in which the rapper takes a stand against men and sticks up for women: “I’m standing up to men who are hairy and stink. You know they think they’re on top, but I grab them by the stank.”

It’s that same distinctive voice, crude humor, crafted lyricism, and cutting insults for which he is famous. The rapper spits out rapid-fire verses, incorporating witty, clever jokes, while using pronunciation and diction as tools to skillfully bend vowels and emphasize phonemes, rhyming words people never thought could be rhymed.

However, this is not the real Slim Shady; it is a synthesized voice imitation whose lyrics and vocal performance are the product of a text-generating algorithm and impressive artificial intelligence (AI) audio speech synthesis. The song amounts to nothing more than a fun gimmick, but the implications of the technology that created the track are profound. Today, anyone with Internet access can download an app that allows them to take existing audio and alter it to make any particular person realistically look like they said something they would never say.

And the same is true for images. Even video can be faked. Take the video, appearing nearly a month after the Russian invasion into Ukraine, of Ukrainian President Volodymyr Zelenskyy. In the minute-long video, he appears behind a presidential podium, the Ukrainian crest of arms emblazoned on the backdrop behind him and informs the public that he

4. Id.
5. Genius, Eminem Proves There Are Plenty of Words That Rhyme with ‘Orange’, YOUTUBE (Dec. 30, 2016), https://www.youtube.com/watch?v=lPcR5RVXHMg [https://perma.cc/ARU8-6HF7].
has made the “difficult decision” to surrender to Russian forces. Staring solemnly into the camera, he speaks directly to his people: “there is no more tomorrow . . . Lay down your arms and return to your families. You should not die in this war.” The message broadcast on live television in Ukraine and appeared on the news station’s website, as well as social media. The video, unsurprisingly, went viral.

But the forgery failed to bring about the hoped-for effect: people were quick to spot some tell-tale clues of deepfakery, and social media platforms quickly removed the fake content. Fortunately, the public had been warned; predicting that the Russians might deploy deepfake technology in the war, the Ukrainian Center for Strategic Communication had released a series of tweets, cautioning its citizens about deepfaked disinformation: “[i]magine seeing Vladimir Zelensky on TV making a surrender statement. You see it, you hear it—so it’s true. But this is not the truth! This is deepfake technology.”

Seeing is believing—but what happens when you trust your own eyes? Deepfake technology—which uses AI to create digitally altered audio, images, and videos that appear legitimate—has rapidly improved and is capable of producing audiovisual imagery that is increasingly difficult to discern from genuine recordings. While there are undoubtedly positive applications of the technology, with the increasing sophistication and accessibility of deepfake technology, novel forms of abuse, exploitation,
and manipulation have emerged.\textsuperscript{15} Among these are face-swapped pornography and fake news.\textsuperscript{16} Deepfake technology, therefore, raises grave concerns for one’s reputation, personal privacy rights, and national security as a whole.

Currently, only three states have laws that directly deal with deepfakes.\textsuperscript{17} Legal scholars have suggested that deepfake transgressions be dealt with under various areas of existing U.S. law such as defamation, copyright infringement, revenge pornography, and harassment.\textsuperscript{18} Though the current legal framework on its face appears sufficient to address deepfake transgressions, in practice, applying existing laws to the problems deepfakes pose fails to adequately address the unique issues presented by the technology. Even the proposed “DEEP FAKE ACCOUNTABILITY ACT”—which would require that all deepfakes contain a watermark and a disclaimer to be legal—vastly mischaracterizes deepfakes as a mere labeling issue and provides no real redress to victims, who would remain responsible for pursuing damages and initiating proceedings against deepfake creators despite being inadequately positioned to do so.

Existing laws—both federal and state, and both civil and criminal—are insufficient to prevent, address, or remedy the harms caused by deepfakes. Further, First Amendment free speech protections provide significant challenges to the government’s ability to regulate deepfakes. In Part I of this Note, I will explore the characteristics of deepfakes that make preventing both the misuse of the technology and its proliferation on social media and the Internet so difficult. Part II will then continue by explaining the unique potential harms that deepfakes pose. Part III will address how the current legal framework fails to adequately prevent these harms or provide remedy to victims, including how Section 230 of the Communications Decency Act serves as a major barrier to redress. In Part IV, I will suggest potential remedies for deepfake transgressions by broadening the language of the Violence Against Women Act Reauthorization Act to include deepfake pornography transgressions, creating state statutes which classify deepfake pornography as a specific type of sexual cybercrime, and educating the public on both the existence of deepfakes and the threats that they pose in order to reduce the potential harm that political deepfakes can inflict.

\textsuperscript{15} See discussion infra Part I.C (discussing the harms of deepfake technology).

\textsuperscript{16} See infra notes 56, 77 and accompanying text.

\textsuperscript{17} See Abigail Loomis, Deepfakes and American Law, Davis Pol. Rev. (Apr. 20, 2022), https://www.davispoliticalreview.com/article/deepfakes-and-american-law#:~:text=The%20only%20states%20with%20legislation,specific%20subset%20of%20informational%20deepfakes \[https://perma.cc/429U-B6FP\] (“The only states with legislation concerning deepfakes are Virginia, Texas, and California.”).

\textsuperscript{18} See discussion infra Parts II.A–II.D.
I. BACKGROUND

A. What Are Deepfakes?

A deepfake is “a video that superimposes hyper-realistic faces onto the bodies of others with the intent of creating a new video with fake representations.” Deepfakes are a kind of AI-generated synthetic media that swaps one person in an image, video, or audio recording with another person’s likeness. Deepfakes take their name from the technology used to generate this fake content: deep learning.

Deep learning is a subset of machine learning technology that uses multiple layers of neural networks—a specific structure of organized algorithms—to process data, discover, and then create patterns. Deep learning stacks these algorithms in a hierarchy of increasing complexity, with each level building from the knowledge gained from the preceding level of complexity. This iterative process means that the longer the algorithm runs, the more it knows.

Like all deep learning computer networks, deepfake technology generates content based on data input: the algorithm is fed large data sets of real recordings or images to effectively learn what a particular face looks like at different angles, in different lightings, and with different expressions. The algorithm then builds an adaptable model of the facial and vocal characteristics of a person which can then be digitally inserted over the face of a different person in a different recording seamlessly— as if it were a mask.

Early, less sophisticated deepfakes triggered an uncanny valley effect—the phenomenon in which a human being experiences a negative emotional response to things that appear somewhat human-like but are clearly not. Since the emergence of generative adversarial networks...
GANs (GANs) in recent years, however, the final outputs of deepfake technology have quickly become more sophisticated and convincing.

GANs use two distinctive neural networks to train its model: a generator and a discriminator. The networks are functionally adversarial and programmed to compete against each other, “mimicking the back-and-forth between a picture forger and an art detective who repeatedly try to outwit one another.” The first network—the generator—generates fake images based on patterns it “learns” from an existing image data set. The second network—the discriminator—learns to identify whether an input image is authentic or computer-generated. The generator’s job is to trick the discriminator into believing the images are real. As content cycles back and forth between the two networks—one algorithm producing the deepfake, the other attempting to detect deepfaked images—both networks continually improve: the discriminator gets better at identifying AI-generated images, and the generator produces more and more realistic images as it attempts to trick the discriminator. The result of this reciprocal feedback loop: the generator can produce fake images with high fidelity, resulting in realistic impersonations that are increasingly indistinguishable from genuine images and recordings.

While the technology behind deepfakes is highly sophisticated, its producers don’t need to be. As the quality of deepfakes has radically improved in recent years, the technology used to create them has simultaneously become easier to access and more affordable. FakeApp, ZAO, and Reface are just some of the face-swapping

---

29. Id.
30. Id.
programs that are available to download on any phone, tablet, or laptop for free. The result is that anyone with access to a YouTube tutorial and enough computing power can produce their own convincing forgery.\footnote{Christopher, supra note 33.}

The increasing availability and affordability of deepfake technology has contributed to its rapid proliferation: over a ten-month period from December 2018 to September 2019, there was an almost one hundred percent increase in the number of deepfake videos online.\footnote{HENRY AIDER ET AL., THE STATE OF DEEPFAKES: LANDSCAPE, THREATS, AND IMPACT 1 (2019).} One piece of software commonly used to create deepfakes was downloaded more than 100,000 times in the first month after being made public.\footnote{Dave Lee, Deepfakes Porn Has Serious Consequences, BBC NEWS (Feb. 3, 2018), https://www.bbc.com/news/technology-42912529 [https://perma.cc/E54H-E8U5].} These programs are specifically designed to allow the average consumer without any programming experience to create their own deepfakes with little effort: simply download the app and feed photos into the program.\footnote{ADAM DODGE ET AL., USING FAKE VIDEO TECHNOLOGY TO PERPETRATE INTIMATE PARTNER ABUSE 5 (2018).}

Still, the quality of the output depends on the volume of information put into the training data set: the software requires hundreds of images and videos to train the program to learn the features of the target object to produce truly convincing fakes.\footnote{THANH THI NGUYEN ET AL., DEEP LEARNING FOR DEEPFAKES CREATION AND DETECTION: A SURVEY 1 (2019).} This is partially why, up until this point, high-profile figures have been the main target for deepfaking, as their public profiles provide plenty of source material for an AI to learn from. However, as Joseph Foley points out, “with the number of selfies the average person takes in a lifetime and rapid technological advances, perhaps soon anyone could be used as a source.”\footnote{Joseph Foley, 14 Deepfake Examples That Terrified and Amused the Internet, CREATIVE BLOQ (Mar. 3, 2022), https://www.creativebloq.com/features/deepfake-examples [https://perma.cc/KC7R-E8KT].} Already, technology has been developed to gather publicly available accounts and websites: “[o]pen-source tools like Instagram Scraper and the Chrome extension DownAlbum make it easy to pull photos from publicly available Facebook or Instagram accounts and download them all onto your hard drive.”\footnote{DODGE ET AL., supra note 40, at 7.} Altogether, by slashing the resources required to produce realistic fabricated content and “democratizing” the process through the dissemination of user-friendly software tools,\footnote{COLLINS, supra note 31.} these tools allow for
“cheap and easy fabrication of content that hijacks one’s identity—voice, face, body.”

B. How Are They Used?

Deepfakes can be used for a variety of purposes, not all of which are harmful. The most obvious beneficial uses of the technology come in the arts and have already been used for that purpose: In *Star Wars: The Last Jedi*, filmmakers used automated dialogue replacement (ADR) technology to fake additional dialogue using snippets from real recordings after Carrie Fisher’s death. A TikTok page dedicated to Tom Cruise deepfakes, which feature the actor showing off magic coin tricks, eating cereal with Paris Hilton, and eating a lollipop, quickly racked up tens of millions of views on the platform. Snapchat filters that superimpose someone’s face on a person’s own in real time, digital avatars, and apps that allow a person to virtually try on online clothes while shopping are other positive uses of the technology.

The technology can also be used to create satirical content that comments on politics and pokes fun at governmental figures. In 2018, video footage of President Obama cursing and calling President Trump a derogatory name appeared online. The video caused a stir, not because of the nature of the remarks, but because the video was entirely fake. Actor and director Jordan Peele had fabricated the video entirely using deepfake technology. He intended for the video to serve as a warning against deepfakes. The public service announcement began by noting,

---


50. BuzzFeedVideo, supra note 25.

51. Brown, supra note 49.


53. Id.
“We’re entering an era in which our enemies can make it look like anyone
is saying anything at any point in time.”

C. What Harms Do They Pose?

The proliferation of increasingly realistic fabricated content presents
numerous potential risks to individuals, organizations, and societies.

1. Threat to Individual Privacy

The first deepfake videos to circulate widely surfaced in 2017 when a
Reddit user posted doctored porn clips featuring the faces of celebrities
such as Emilia Clarke, Taylor Swift, Scarlett Johansson, and Gal Gadot,
among others, swapped onto the faces of porn performers. Pornography
continues to account for the vast majority of deepfakes: a 2019 study
found that, of the 14,678 deepfake videos on the Internet, 96% were
pornographic. While the majority of pornographic deepfakes feature
actresses and musicians working in the entertainment industry, everyday
citizens have been victimized by deepfake pornography as well.

Nonconsensual deepfake pornography is essentially the next iteration
of revenge pornography, representing a new and degrading means of
humiliation, harassment, and abuse. While revenge pornography involves
leaking a real nude image or video initially shared privately, deepfake
pornography allows the perpetrator to fabricate a pornographic video
starring any woman who has shared images of herself on the Internet. Most disturbingly, because these videos are built off public photos,
anyone can be a victim. After being superimposed into dozens of graphic
sex scenes (including one video that was falsely described as real
“leaked” footage and has been watched on a major porn site more than
1.5 million times), Scarlett Johansson stated, “Nothing can stop someone
from cutting and pasting my image or anyone else’s onto a different
body.” While the creation of deepfakes requires some technical know-
how, a marketplace sprung up for people seeking these videos.

54. BuzzFeedVideo, supra note 25.
55. Ian Sample, What Are Deepfakes – and How Can You Spot Them?, GUARDIAN (Jan. 13,
2020, 5:00 PM), https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-
how-can-you-spot-them [https://perma.cc/LA99-3AJD].
56. AJDER ET AL., supra note 38, at 6.
57. See id. at 2 (stating that all but one percent of the subjects featured in deepfake
pornography videos were actresses and musicians working in the entertainment sector).
58. Anne Pechenik Gieseke, “The New Weapon of Choice”: Law’s Current Inability to
Properly Address Deepfake Pornography, 73 VAND. L. REV. 1479, 1481 (2020).
59. Drew Harwell, Fake-Porn Videos Are Being Weaponized to Harass and Humiliate
post.com/technology/2018/12/30/fake-porn-videos-are-being-weaponized-harass-humiliate
women-everybody-is-potential-target/ [https://perma.cc/33WS-PK4L].
60. Gieseke, supra note 58, at 1485; Romano, supra note 52.
Deepfake producers offer to create videos by request on forum-based websites like 4chan, 8chan, and Voat. What is the going rate for these videos? About twenty dollars per fake.

Like revenge porn, deepfake pornography can be used as a powerful instrument of individual intimidation, coercion, or defamation and can cause problems in the context of intimate partner abuse. “A fake video that causes an audience to believe that a partner was featured in revenge porn can cause precisely the same kind of reputation, privacy, and property harms, and can rob people of their potential for the rest of their lives.” While the video may not be real, the psychological damage to the individual is. Because, while the sex scenes look realistic, they are not consensual cyber porn. “Face-swapped porn inflicts the harm of sexual objectification without consent. Like nonconsensual porn, face-swapped porn violates the partner’s expectation that all aspects of sexual activity should be founded on consent.”

In addition to the psychological effects of being the subject of nonconsensual deepfake pornography, compromising images and videos can damage a victim’s reputation, rendering them “unemployable, undateable, and potentially at physical risk.” A single intimate image can quickly dominate the first several pages of search engine results for the victim’s name, meaning that one Google search could uncover a deepfake sex tape in which the victim did not participate, permanently affecting her ability to find a job. “Deepfake technology is being weaponized against women.”

2. Threat to National Security and Politics

Perhaps even more disturbing than the harm to individuals that deepfakes can cause are the national security implications of sophisticated forgeries: a well-executed and well-timed deepfake has the potential to cause significantly destabilizing political impacts.

In May 2019, a video of U.S. Speaker of the House and Democrat Congresswoman Nancy Pelosi went viral on social media. In the video, Pelosi’s speech appeared slurred, as she stumbled over her words. The video “was retweeted by the official Twitter account of U.S. President Trump, receiving over 6.3m views” in the three months after it was posted. “On a popular Facebook page, the video received over 2.2m views.”
views in the forty-eight hours following its initial upload, with commenters calling Pelosi ‘drunk’ and a ‘babbling mess.’” 69

However, the video wasn’t real. “Pelosi’s speech had been slowed down” to make her appear drunk. 70 While the video was ultimately harmless, the rapid spread of this manipulated video 71 demonstrates the potential for deepfakes to generate significant harm on a much broader scale—creating social unrest, political distrust, and delegitimizing the news media.

Crude deepfakes—such as the Zelenskyy deepfake mentioned earlier—have already been created and spread by foreign powers in attempt to disrupt democracy by targeting political leaders. For example, in 2009, John Beyrle, former U.S. Ambassador to Moscow, was the subject of a fake sex video which appeared to show him having an affair. 72 The video was supposedly disseminated by the Russian government and was created to harm his reputation. 73 Similarly, Michael McFaul, who served as the American ambassador to Russia from 2012 to 2014, was accused of pedophilia, having had his face inserted into photographs and his speech “spliced . . . to make me say things I never uttered.” 74

Altogether, the technology opens up the potential that any foreign government could use the technology for any means—including depicting an American politician using a racial epithet, taking a bribe, or encouraging certain political action. 75 In 2018, a deepfake video of President Trump was released where he appears to call on Belgium to exit the Paris Climate Agreement. 76 In the fictional address, he is shown saying: “As you know I had the balls to withdraw from the Paris climate agreement. And so should you.” 77 A Belgian political party created the fake video, apparently to “start a public debate,” but it is easy to see how the technology could easily be leveraged for more nefarious means. 78

69. Id.
70. Id.
71. Scholars refer to videos like these that have been manipulated with basic editing tools or intentionally placed out of context as “shallowfakes.” See id.
73. Id.
74. Id.
75. Id.
77. Id.
78. Id.
Senator Marco Rubio has spoken out against the harms deepfakes pose, noting that “it’s a weapon that could be used—timed appropriately and placed appropriately—in the same way fake news is used, except in a video form, which could create real chaos and instability on the eve of an election or a major decision of any sort.”

The spread of deepfakes will threaten to erode the trust necessary for democracy to function effectively, for two reasons. First, and most obviously, the marketplace of ideas will be injected with a particularly dangerous form of falsehood, as deepfaked videos—though false—purport to be truth. As journalist Franklin Foer explains in The Atlantic, what makes deepfakes so frightening is “the acuity of the technology: A casual observer can’t easily detect the hoax.” Second, and more subtly, “the public may become more willing to disbelieve true but uncomfortable facts.”

As Senator Rubio warned, “deepfakes pose an especially grave threat to the public’s trust in the information it consumes; particularly images, and video and audio recordings posted online.” There is a potential for weaponization of the idea that we cannot believe any image could be wielded by authoritarians and totalitarians worldwide. News organizations may hesitate from rapidly reporting real, disturbing events for fear that the evidence of them will turn out to be fake.

II. WHY CURRENT LEGAL FRAMEWORKS FAIL

While the existing legal framework provides victims who become the subject of unwanted deepfakes some avenues for redress, those existing legal claims only apply in very specific circumstances, and thus fail to serve as feasible solutions. Tort and copyright law could provide a cause of action for some victims, but both are vulnerable to a First Amendment defense that deepfakes are protected speech. Further, Section 230 of the Communications Decency Act (CDA), which provides substantial protection to online platforms, acts as a powerful barrier to recovery for victims in most circumstances.

79. Riechmann, supra note 72.
81. Chesney & Citron, supra note 45.
84. Id.
A. Tort Law

State tort law typically supplies the remedy for civil privacy violations. Defamation law appears specifically applicable to the threat presented by deepfakes, given that deepfake technology provides the opportunity for anyone’s image to be used in a variety of ways and therefore the potential to cause significant damage to a person’s reputation. While state law governs defamation causes of action and thus the standard varies, 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.\(^{85}\)

Defamatory statements are those that “tend to damage another’s reputation to the extent of lowering their regard in the community or deterring others from associating with them.”\(^{86}\) Images or videos can constitute defamatory “statements,” even if the image is doctored.\(^{87}\)

However, to create liability for defamation there must be publication of matter that is both defamatory and false.\(^{88}\) When deepfakes can be compared to an original recording, proving that the deepfake is in fact fake will be relatively straightforward. However, as the quality of deepfakes improves, proving that a recording is false may require expensive and complex technology.\(^{89}\)

Further, a deepfake cannot constitute “defamation” if the content does not claim to be “real.”\(^{90}\) Disclaimers that the content is a deepfake will likely preclude any liability for defamation; deepfakes, by definition, are doctored, and therefore cannot reasonably be interpreted as stating actual facts about the person involved.\(^{91}\)

Further, a deepfake creator has a powerful defense to any defamation claim by arguing that the video is a parody. In the landmark Supreme

\(^{85}\) Restatement (Second) of Torts § 558 (Am. L. Inst. 1977).

\(^{86}\) Id.

\(^{87}\) See Crump v. Beckley Newspapers, Inc., 320 S.E.2d 70, 80 (W. Va. 1983) (“[I]t is well established that although libel is generally perpetrated by written communication, it also includes defamation through the publication of pictures or photographs.”); Kiesau v. Bantz, 686 N.W.2d 164, 178 (Iowa 2004) (finding that a doctored photo can be defamatory), overruled in part by Alcala v. Marriott Int’l, Inc., 880 N.W.2d 699 (Iowa 2016).

\(^{88}\) Restatement (Second) of Torts § 558 (Am. L. Inst. 1977).

\(^{89}\) See Jason Haas, Deepfake Dilemma, 2019 Intell. Prop. Mag. 33, 33 (“Even expensive discovery measures may prove inadequate to identify a deepfake creator, leaving a plaintiff’s only possible recourse to sue republishers.”).


\(^{91}\) See Hustler Magazine v. Falwell, 485 U.S. 46, 57 (1988) (finding that the magazine’s parody ad could not “reasonably be understood as describing actual facts about [respondent] or actual events in which [he] participated”).
Court case Hustler v. Falwell, televangelist Jerry Falwell sued Hustler magazine for defamation and intentional infliction of emotional distress after it published a satirical ad suggesting he had drunken sex with his mother in an outhouse. The ad contained a disclaimer: “ad parody—not to be taken seriously.”

The Court ruled against Falwell, finding that the ad parody was not believable and therefore did not contain false statements of fact; as a result, the magazine was constitutionally immune from defamation liability. As Erik Gerstner explains,

Defamation is by its nature mutually exclusive of parody. By definition, defamation requires a false statement of fact; parody, to the degree that it is perceived as parody by its intended audience, conveys the message that it is not the original and, therefore, cannot constitute a false statement of fact.

Deepfake videos are also likely to trigger the common law tort of intentional infliction of emotional distress (IIED). Unlike with a defamation claim, the victim of IIED does not need to prove that a statement is false: the only concern with regard to IIED is whether the conduct was “patently offensive”—false or not. However, Falwell further demonstrates that the First Amendment defense extends to other theories of tort liability that may offer individuals redress, such as IIED. The Supreme Court rejected both claims, holding that the First Amendment prohibits public figures from recovering damages for the tort of IIED if the emotional distress was caused by a parody that a reasonable person would not have interpreted as factual. Further, the intent requirement for a claim of IIED is a powerful barrier to remedy for victims of deepfake media: “[w]hile there will clearly be intent in the creation of the media itself, in many cases it is unlikely that a court will find actual intent to cause emotional distress.” Given these constitutional limits, a deepfake can likely only be prohibited if it falsely depicts an individual, “does not include a disclaimer, and is made with

---

92. Id. at 52.
93. Id.
95. Gerstner, supra note 46, at 1, 5.
96. Id. at 5–6.
97. Id. at 6.
100. Gerstner, supra note 46, at 6.
knowledge that it was false or with reckless disregard of whether it was false or not.  

B. Copyright

Copyright law also fails to provide a remedy for those victimized by deepfakes because deepfake creators can likely receive First Amendment protection under the concept of “transformative use.” In April 2020, the YouTube creator Vocal Synthesis, an anonymous YouTube account that uses AI to make vocal impersonations of celebrities and politicians rapping various content, received the first copyright claim for deepfaked audio content. The videos in dispute were two AI-powered voice impersonations depicting Jay-Z rapping William Shakespeare’s “To Be or Not To Be” soliloquy from Hamlet, and Billy Joel’s “We Didn’t Start the Fire.” Like the Eminem video, the videos in question are “entirely computer-generated using a text-to-speech model trained on the speech patterns of Jay-Z.”

Apparently unamused by the Shakespearean jest, Jay-Z’s legal team sought to remove the videos from the platform, claiming in their DMCA takedown notice that the content “unlawfully uses an AI to impersonate our client’s voice.” YouTube initially removed the videos, but ultimately, Jay-Z’s claim was unsuccessful; within a matter of days, YouTube rejected the copyright claims, citing insufficient grounds from the claimant. The videos were promptly reinstated.


103. Id.


105. “DMCA” stands for the “Digital Millennium Copyright Act.” A DMCA takedown notice “informs a company, web host, search engine, or internet service provider that they are hosting or linking to material that infringes on a copyright. The party that receives the notice should take down the material in question as soon as possible.” DMCA Notice: Everything You Need to Know, UPCOUNSEL (Oct. 5, 2020), https://www.upcounsel.com/dmca-notice [https://perma.cc/A9BR-SL2V].

106. Baio, supra note 102.


108. Id.
Though Roc Nation has not yet taken any of their claims to court, the Jay-Z situation and commentary surrounding the initial takedown attempt demonstrate that copyright law is not a viable source of redress for those who unwillingly become the subject of deepfake creations in most circumstances. First, a person’s “sound” or general appearance is not copyrightable; the person depicted in the deepfake must own the source material used to create the deepfake to bring a copyright infringement claim. Second, the doctrine of fair use likely provides a viable First Amendment defense to deepfake creators in most cases, since deepfakes almost by definition qualify as “transformative use.”

Copyright protection only applies to “original works of authorship fixed in any tangible medium of expression.” While 17 U.S.C. § 102 provides copyright protection to “sound recordings” and “musical works,” this protection applies to the underlying musical elements and lyrics of a work. A voice or vocal style, on the other hand, cannot be copyrighted, as those sounds are not “fixed”; rather, there are an infinite number of words or phrases a person could potentially utter in their distinctive voice. Using AI to impersonate someone’s voice, therefore, does not violate existing copyright law.

Similarly, an individual’s appearance is not protectable under copyright because an appearance is “not created like a work of authorship—it simply exists.” Any photographs or videos used to create a deepfake are subject to copyright protection, but existing regulations exclusively cover the person who created the content: the “author.” If the subject of the deepfake did not take and post the source material of the deepfake themselves, he or she has no claim for copyright infringement.

Even when an individual depicted in the deepfake owns all images and video used to create a deepfake, the doctrine of fair use—which

110. Id.
111. See Midler v. Ford Motor Co., 849 F.2d 460, 462 (9th Cir. 1988) (“A voice is not copyrightable. The sounds are not ‘fixed.’”); Butler v. Target Corp., 323 F. Supp. 2d 1052, 1055 (C.D. Cal. 2004) (explaining that although lyrics to a song are copyrightable, the underlying voice is not).
112. Hogan, supra note 104.
113. Zachary Schapiro, Deep Fakes Accountability Act: Overbroad and Ineffective, 2020 B.C. INTELL. PROP. & TECH. F. 1, 13; see 17 U.S.C. § 102 (granting copyright protections to works of authorship); see also Downing v. Abercrombie & Fitch, 265 F.3d 994, 1004 (9th Cir. 2001) (noting that a person’s likeness is not a “work of authorship” within the meaning of the Copyright Act and thus is not subject to copyright protection).
114. The copyright owner is the one who creates the content. See 17 U.S.C. §§ 102, 106(3) (establishing that “pictorial, graphic . . . [and] motion pictures and other audiovisual works” may be protected by copyright by the authors of the work).
115. In cases where a photographer or videographer took the content, ownership may be stipulated in the contract with the photographer or videographer. See Schapiro, supra note 113.
allows for some unlicensed use of material that would otherwise be copyright protected—likely protects most deepfake creators from liability for infringement. Codified in federal law as 17 U.S.C. § 107, “fair use” is a defense based in the First Amendment that allows an infringer to use the original author’s work without asking permission in certain, limited circumstances. These include criticism, comment, news reporting, teaching, scholarship, and research. Parody may also claim fair use under 17 U.S.C. § 107.

To determine whether a particular unlicensed use of a copyrighted work qualifies as “fair,” courts consider four factors: (1) the purpose and character of the creator’s use; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion taken; and (4) the effect upon the use upon the potential market. In the 1994 case *Campbell v. Acuff-Rose Music*, the Supreme Court honed in specifically on the first of the four fair use factors—the purpose and character of the use—and emphasized that the important aspect of the analysis was whether the purpose and the character of the use was “transformative.”

Whether a work qualifies as “transformative” depends largely on whether the work builds upon a copyrighted work in a different manner or for a different purpose from the original—i.e., it transforms or modifies the original work in some creative way so that it creates content with new expression, meaning, or message—or whether it merely copies from the original. However, deepfakes—which take a photo or video and transform it into something vastly different from the purpose or character of the original work—may be the epitome of transformative use. Further, the typical deepfake is not likely meant to replace the original work. Copyright law, therefore, is not a promising source of redress for victims of nonconsensual deepfake content in most circumstances.

### C. Nonconsensual Pornography Laws

While the term “nonconsensual pornography” is defined as the nonconsensual sharing of intimate imagery and encompasses both deepfake pornography and revenge pornography, most nonconsensual...
pornography statutes are too narrowly drawn to provide an effective remedy for individuals depicted in nonconsensual pornographic deepfakes and instead only address revenge pornography offenses.\textsuperscript{124} Revenge pornography typically occurs when an individual is seeking revenge against a former intimate partner by “sharing sexually explicit images that the individual obtained during the period of their intimacy,” even if the image was originally taken with the subject’s consent.\textsuperscript{125}

The majority of states have enacted laws to address the growing epidemic of nonconsensual pornography.\textsuperscript{126} However, the language of most of these laws prevents them from applying to deepfakes, as many state laws against nonconsensual pornography require that the perpetrator had an intent to harm the subject of the images or video to be liable.\textsuperscript{127} For example, Arizona’s nonconsensual pornography statute makes it illegal to disclose an image of a person depicted in a state of nudity or engaged in specific sexual activities only if the image was disclosed with the intent to “harm, harass, intimidate, threaten or coerce” the depicted person.\textsuperscript{128} Similarly, Colorado law makes “[p]osting a private image for harassment” a misdemeanor offense where a person posts any photograph or video that displays the private intimate parts of another person “with the intent to harass the depicted person and inflict serious emotional distress upon the depicted person” and where “the conduct results in serious emotional distress of the depicted person.”\textsuperscript{129} These intent-to-harm requirements are problematic for victims of deepfake pornography because deepfake creators may not intend to hurt their subject; perpetrators may be motivated by voyeurism, profit, or a variety of other

\textsuperscript{124.} Nonconsensual Pornography (Revenge Porn) Laws in the United States, BALLOTPE\textsuperscript{125.} supra note 98, at 224.


127. See OHIO REV. CODE ANN. § 2917.211(B)(5) (West 2022) (“No person shall knowingly disseminate an image of another person if . . . The image is disseminated with intent to harm the person in the image.”); MO. REV. STAT. § 573.110.2.(1) (2018) (requiring that a person must “intentionally disseminate” sexual imagery “with the intent to harass, threaten, or coerce an image of another person”); WASH. REV. CODE § 9A.86.010(2)(a) (2015) (“A person who is under the age of eighteen is not guilty of the crime of disclosing intimate images unless the person: (a) Intentionally and maliciously disclosed an intimate image of another person.”); OR. REV. STAT. § 163.472(1)(a) (2021) (“The person, with the intent to harass, humiliate or injure another person, knowingly causes to be disclosed an identifiable image of the other person whose intimate parts are visible or who is engaged in sexual conduct.”).


129. COLO. REV. STAT. § 18-7-107 (2012).
reasons. Yet deepfake creators that share their content online without any harmful intent would not be liable in states with these intent-to-harm requirements.

Further, many nonconsensual pornography laws also require the subject to have had a reasonable expectation of privacy in regard to the content for the distributor to be liable. That is because the basis for these laws is that what is being shared is private, true information that is being disclosed without the subject’s consent. However, deepfakes are generally produced using photographs the victims themselves have posted online. The deepfakes themselves, therefore, are not legally in violation of privacy; the victim had no reasonable expectation of privacy in regard to the source material they posted.

Additionally, many state statutes have language that specifies that the reasonable expectation of privacy occurred in regard to the taking of the nude images or videos. For example, Tennessee’s nonconsensual pornography law requires that the “image was photographed or recorded under circumstances where the parties agreed or understood that the image would remain private.” Sharing a deepfake would not satisfy this law because a deepfake is neither photographed nor recorded.

Federal law remedies for nonconsensual pornography have similarly limited application when considering pornographic deepfakes. In March 2021, President Biden signed into law the Violence Against Women Act Reauthorization Act of 2022 (VAWA), which, in part, established a federal civil cause of action for individuals whose intimate visual images are disclosed without their consent and allows a victim to recover damages and legal fees. However, a plaintiff must prove that


131. See MO. REV. STAT. § 573.110.2.(2) (2018) (“A person commits the offense of nonconsensual dissemination of private sexual images if he or she: . . . (2) Obtains the image under circumstances in which a reasonable person would know or understand that the image was to remain private.”); WASH. REV. CODE § 9A.86.010(1)(a) (2015) (“A person commits the crime of disclosing intimate images when . . . the person disclosing the image: (a) Obtained it under circumstances in which a reasonable person would know or understand that the image was to remain private.”); D.C. CODE § 22-3052(a)(2) (2015) (“here was an agreement or understanding between the person depicted and the person disclosing that the sexual image would not be disclosed”).

132. Dold, supra note 90.

133. Gieseke, supra note 58, at 1501–02.

134. Id.


defendant was aware of a substantial risk that the person depicted in an image expected it would remain private and that they did not give consent to its distribution.\footnote{137}

Currently, only four states have laws specific to deepfaked pornographic media.\footnote{138} Georgia’s prohibition on nude or sexually explicit electronic transmissions makes it a criminal offense to share or post images or videos of “nudity or sexually explicit conduct of an adult, including a falsely created videographic or still image” without the consent of the depicted person.\footnote{139} Virginia amended its nonconsensual pornography statute to include “a person whose image was used in creating, adapting, or modifying a videographic or still image with the intent to depict an actual person and who is recognizable as an actual person by the person’s face, likeness, or other distinguishing characteristic.”\footnote{140} In October 2019, California Governor Gavin Newsom signed AB 602 into state law, which provides a private cause of action for individuals to sue creators of deepfake pornography.\footnote{141} California defines “depicted individual” as any individual “who appears, as a result of digitization, to be giving a performance they did not actually perform or to be performing in an altered depiction.”\footnote{142} Finally, New York’s nonconsensual pornography statute, which creates a private right of action for unlawful dissemination or publication of a sexually explicit depiction of an individual defines “sexually explicit material” as “any portion of an audio visual work that shows the depicted individual performing in the nude,” making the language broad enough to accommodate deepfake pornography claims.\footnote{143}

D. \textit{Section 230 of the Communications Decency Act}

Another obstacle to redress for victims of deepfakes is Section 230 of the Communications Decency Act, which protects Internet service

\footnote{137. VICTORIA L. KILLION, CONG. RSCH. SERV., LSB10723, FEDERAL CIVIL ACTION FOR DISCLOSURE OF INTIMATE IMAGES: FREE SPEECH CONSIDERATIONS 2 (2022).}


\footnote{139. GA. CODE ANN. § 16-11-90(b)(2) (2021) (emphasis added).}

\footnote{140. VA. CODE ANN. § 18.2-386.2A. (2014).}


\footnote{142. CAL. CIV. CODE § 1708.86(a)(4) (West 2020).}

\footnote{143. N.Y. CIV. RIGHTS LAW § 52-c(e) (McKinney 2021).}
providers from liability for content published by users on their portals.144 Deepfakes come, in many cases, with an attribution problem: technologies may be employed to allow the creator to remain anonymous, such as disconnecting IP addresses from the post.145 Identifying the creator of a deepfake, therefore, may be impossible.146 Absent an identifiable defendant, a plaintiff may only be able to pursue the disseminator of the deepfake—the host website or social media platform.147 Yet under Section 230, host websites may not be held liable for publication of pictures or videos posted by a third party or any damage it causes—whether illegal or not—which leaves only the producer of the deepfake potentially liable for the harm.148 Because of this, Section 230 essentially leaves victims with no practical recourse. However, it is worth noting that the scope of the act does not cover intellectual property breaches, so if a party holds copyright to the image, the takedown request must be executed pursuant to the Digital Millennium Copyright Act.149

Repealing Section 230, however, would likely raise constitutional First Amendment issues. Instead, legal expert Mary Anne Franks, president of the Cyber Civil Rights Initiative and a professor at the University of Miami School of Law, has argued that federal identity theft law should be amended to frame social media users as the consumers of content and therefore invoke consumer protection rights.150 Doing so would place the distribution of deepfake content alongside the misappropriation of information such as names, addresses, or social security numbers, and would serve as a powerful deterrent against the distribution of malicious deepfakes.151

144. See 47 U.S.C. § 230(c)(1) (providing that “[n]o provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider”).
146. See Gieseke, supra note 58, at 1495 (“Producers of deepfake pornography can simply vanish from the internet—or take precautions to ensure that they cannot be tracked down.”).
147. Reid, supra note 98, at 218.
148. Gieseke, supra note 58, at 1494.
150. Mutale Nkonde, Congress Must Act on Regulating Deepfakes, MEDIUM (June 17, 2019), https://onezero.medium.com/congress-must-act-on-regulating-deepfakes-1e7e94783be8 [https://perma.cc/29M4-CC6Z].
151. Id.
E. DEEP FAKES Accountability Act

On April 8, 2021, Representative Yvette Clarke (D-NY) introduced the DEEP FAKES Accountability Act into Congress. The bill, whose acronym stands for “Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act,” aims “[t]o combat the spread of disinformation through restrictions on deep-fake video alteration technology.”

In its attempt to limit the potential damage of synthetic media which appears to be authentic, the DEEP FAKES Act would require anyone creating a piece of fabricated media that imitates a person to disclose that the video is altered or generated using AI, by placing “embedded digital watermarks” on the content, as well as textual descriptions accompanying the image or video. If the altered content contains audio, the piece must also include a “clearly articulated verbal statement that identifies the record as containing altered audio and visual elements, and a concise description of the extent of such alteration.” Failing to do so would be a crime.

The proposed Act would also establish a private cause of action for victims of deepfaked media, which it terms “advanced technological false personation,” to sue the creators and vindicate their reputations in court. The bill defines “advanced technological false personation” broadly, providing that “the word ‘advanced’ within the term advanced technological false personation record shall not be interpreted as narrowing the definition of such term,” in order to apply to new technologies as they advance. In order to protect a victim’s privacy, these documents may be filed under seal “if such plaintiff can demonstrate a reasonable likelihood that the creation of public records regarding the advanced technological false personation record would result in embarrassing or otherwise harmful publicization of the falsified material activity in an advanced technological false personation record.”

Even with these safeguards, however, the proposed act fails to adequately protect victims of deepfaked content, as malicious actors can
remain anonymous while easily circumventing the Act’s requirements. Watermark and other metadata-based markers are usually trivial to remove; text can be cropped, logos removed, and even a sophisticated whole-frame watermark can be eliminated simply by being re-encoded for distribution on a different platform.\footnote{Devin Coldewey, \textit{DEEPFAKES Accountability Act Would Impose Unenforceable Rules—But It’s a Start}, TECHCRUNCH (June 13, 2019, 3:25 PM), https://techcrunch.com/2019/06/13/deepfakes-accountability-act-would-impose-unenforceable-rules-but-its-a-start/ [https://perma.cc/R2W6-JJPC].} Though the Act included the creation of several task forces and coordinators to provide victim assistance, it includes no real enforcement mechanism. Though the Act would create a task force at the Department of Homeland Security that would lead the charge against combatting deepfakes, the taskforce would serve more of a research and reporting function, as well as collaborating with private sector companies such as social media platforms in their attempts to prevent malicious deepfakes.\footnote{Id.} Altogether, the Act conceives of deepfakes as a labeling issue, whereas real prevention and redress would require measures far more tailored to the harm the technology poses.

III. RECOMMENDATIONS

Based on the earlier mentioned issues surrounding deepfakes and the existing legal remedies, the federal government should amend the Violence Against Women Act Reauthorization Act with a broad enough definition to encapsulate nonconsensual deepfake pornography. Additionally, the federal government should provide a definition of deepfakes. Even if no federal deepfake definition is ultimately provided, states should create their own deepfake laws. In order for state nonconsensual pornography laws to apply to victims of deepfake pornography too, states must amend their laws to remove intent-to-harm and reasonable expectation of privacy requirements and expand the definition to include deepfake depictions.

CONCLUSION

Regulation of deepfakes is essential due to the negative consequences that could arise on both the individual and national levels from the malicious use of the technology. A multi-faceted approach to combating malicious deepfakes is necessary, and an effective one would include amending federal and state legislation, as well as coordination with social media networks and Internet companies. The law often lags behind technology, and until proper legal measures are put into place, the best strategy to combatting deepfake harms may be education. The Zelenskyy
deepfake gained no traction, in part, because the public had been warned. A knowledgeable public who is aware of the deepfake phenomena and looks at media with a questioning mind may be the best tool to combat the harms that deepfakes can inflict.