Intersect, Vol 14, No 3 (2021)

Artificial Illusions: Deepfakes as Speech

Nicolas Graber-Mitchell
Amherst College

Deepfakes, a new type of artificial media created by sophisticated machine learning algorithms, present a fundamental epistemological problem to society: How can we know the truth when seeing and hearing are not believing? This paper discusses how deepfakes fit into the category of illusory speech, what they do in society, and how to deal with them. Illusions present an alternate reality, much like a lie, but they also contain evidence for that reality. Some illusions, like games of tag and magic tricks, are harmless and fun. Others, like counterfeit coins and deepfakes, harm others and are deeply convincing. For example, the most common use for deepfake technology is to produce pornographic videos of women who never consented. After strangers attacked them in this way, women reported feeling violated and living in a state of constant "visceral fear." Pornographic deepfakes — most often deployed against women — abridge their targets' sexual agency and privacy, contributing to inequality and enabling intimate partner abuse, workplace sexual harassment, and other discrimination and hate. Deepfakes also pose a threat in politics and society more generally. In addition to allowing malicious actors to produce convincing, illusory disinformation, their increased use may lead to a general inability to discern the truth. In the face of the deep and distressing harms that deepfakers cause to women and the danger that they present to democracy, this paper argues for new civil and criminal penalties for deepfakers as well as new regulations and liabilities for internet platforms that host their work.

Introduction

In December 2017, a staff writer at Vice discovered a pornographic video of Gal Gadot, an Israeli actress of *Wonder Woman* fame, having sex with her supposed stepbrother (Cole, 2017). A user with the screenname 'deepfakes' created and posted the video on the link-sharing website Reddit along with other videos of other celebrities having sex. The only problem is that Gal Gadot never participated in a pornographic film.

Instead, what the reporter at Vice found was the product of a new type of artificial intelligence that swaps one person's face with another in a video. Depending on the sophistication of the artificial intelligence one uses, the resulting videos can mimic details as tiny as mouth movements and facial expressions — and appear disturbingly realistic. The original Reddit user ‘deepfakes’ did not employ teams of CGI artists, huge banks of computers, or even cutting-edge technology to produce his illusory pornography. Instead, he used nothing but a consumer-grade computer, publicly available software, and digital images of his targets.

Eventually, the manufactured videos themselves became known as “deepfakes.” As deepfake technology spread around the internet, other people discovered new uses for it, like generating audio that mimics the speech of famous politicians (Gholipour, 2017) and creating illusory videos of President Obama (Choi, 2017).

Since deepfakes are illusions, and present an alternate reality where people do things they never did and say things they never said, they pose a unique problem to our society and inflict unique harms on individuals. In pornography, deepfakes are essentially nonconsensual sex, and in politics, deepfakes make it impossible for voters to discern the truth, among several other harms. When presented with these troubling videos, what should we do to minimize and eliminate their harms? Since deepfakes are speech, what does the First Amendment prevent government from requiring of deepfake creators and internet platforms? In the rest of this essay, I will explain what deepfakes are and why they are so effective, how they act on the world, the specific harms they cause, and what we can do about it.

What is a deepfake?

Deepfakes are videos, audio, or images of human beings generated, either in part or totally, by advanced artificial intelligence networks. Unlike other types of edited, manipulated media, like airbrushed or photoshopped pictures, deepfakes do not require huge amounts of careful human effort. While deepfaking is a skill, and one’s deepfakes become more and more convincing with practice, even laypeople can create them.

Deepfakes take advantage of deep learning, a field of artificial intelligence that has grown in the last decades. Deep learning algorithms consist of a network of linked “neurons” that each receive many inputs in the form of numbers, process those inputs according to a randomly generated weight function, and then output a single number that other neurons in the network use as input. To “train” the algorithm, users feed the network huge amounts of training data along with expected results. The network processes the data and compares its output with the expected results to tweak the weight functions of its neurons. Eventually, after processing enough training data, the deep learning network — also known as a neural network — can successfully process data similar to, but not in, its training set.

In 2014, a team of machine learning experts developed a new approach to machine learning known as the generative adversarial network (GAN) (Pan et al., 2019). To produce better results, GANs pair a generating neural network with a “discriminator,” an entirely separate network designed to detect bad results. Then, the generator and the discriminator learn together. They compete against each other until the generator finally produces data so true to life that the discriminator cannot tell it from the real thing. These networks proved extraordinarily effective at producing realistic images, text, and other data from a training corpus.

Deepfakers use GANs trained on hundreds of images of a target person to generate new material all-but indistinguishable from the real thing (Korshunov & Marcel, 2018). The resulting video deepfakes appear so realistic because the network that generated them spent hours refining its internal weights to evade detection by its partner network, the discriminator. In essence, the discriminator fulfills the role of a human being, since we can easily discern fake faces from real ones with nothing more than a glance (Lewis & Edmonds, 2003). Once the discriminator, and humans by extension, can no longer tell the deepfaked results from the real thing, the deepfaker is done. These algorithms do not only work on images: GANs can also generate deepfaked audio if given a training set of human audio samples.

Though GANs only emerged in 2014, the technology itself is open-source and widely available. Industry-grade neural network frameworks developed by independent programmers and major companies are free for download and use on the internet. Using these basic tools, it is easy for even hobbyist programmers to dabble in machine learning. Putting together a reasonably effective neural network is as simple as placing basic building blocks in sequence. Though sophisticated results like deepfakes require more tuning and expertise, the first deepfaker turned his hobbyist dabbling into an open-source project with dozens of contributors. Anyone, with a few clicks, rudimentary knowledge of running command-line programs, and enough training images, can download the project, run the code, and generate deepfakes.

More established and knowledgeable actors can easily roll their own software and utilize GANs to create more realistic illusory media of any type. Moreover, as deepfake detection metastasizes into its own field of study, deepfake technology grows ever more effective since its creators can learn from advances in detection to supercharge their adversarial learning. Right now, detection techniques involve specific, current deficiencies of deepfake technology. For example, deepfaked faces rarely blink with typical human patterns. However, future deepfakers will almost certainly incorporate the available knowledge on deepfake detection into their models. Another common detection technique involves matching the surrounding video, which is relatively unaltered when deepfakers swap faces, with videos that already exist. If a match is found, that provides strong evidence that the later video is deepfaked. However, anyone with a slightly higher budget who can afford to hire actors specifically for the deepfake can evade this style of detection.

Deepfake detection and creation, much like the underlying networks used to make deepfakes, will always be locked in an escalating arms race. Though the most advanced deepfakes right now can only generate human faces and place them into already-recorded surrounding video, it is not difficult to imagine near-future GANs that build entire scenes, including personalized human body movements, out of whole cloth. As more time passes, deepfake technology and its potential for illusion only grows.

Illusion

Illusion is the act of illustrating the world contrary to reality. In this way, deepfakes are illusion. They portray something as occurring in real-life when, in fact, it never occurred. When a magician pulls a coin from behind your ear, they do the same — they seek to convince you, through the absolute evidence of the coin's existence, that it came from behind your ear. Likewise, optical illusions convince you that you are seeing something that does not exist, such as movement on a static page.

Falling for an illusion leaves one misled or deceived, either convinced of an untruth or unsure as to the actuality of what they have perceived. When we misapprehend after seeing an illusion, we do not necessarily fail to understand the real world. Rather, we may perceive and understand an alternate world, one that differs from the real world in at least some ways. Depending on how much we hold onto our preconceptions of a non-illusory world, we may doubt illusion, but the hallmark of a good illusion is its ability to introduce even a tiny amount of disbelief into what we thought was true before. After the magician makes a volunteer disappear, they force their audience to grapple with the possibility that the volunteer actually *did* disappear through the arcane, mysterious forces of magic rather than a clever trick of attention.

In fact, illusion goes hand in hand with the suspension of disbelief. Only someone who wants to be bored enters a magic show with the object of doubting the illusions onstage and never giving themselves over to amazement. Likewise, we do not settle into movie theater seats only to gripe about how the computer-generated graphics of monsters and aliens do not accurately portray reality. We enter the theater with the full knowledge that for the next two hours, our enjoyment hinges on believing the illusions beaming out of the screen.

Illusion, writ large, is not nefarious or evil. Instead, illusion is fun.[1] Stories, whether told through illusory graphics or not, play off our ability to fall into a world of untruth and fiction, forgetting our surroundings and losing ourselves within an alternate reality. When we convince ourselves that the floor is lava and jump from couch to table in order to avoid burning in the molten rock, we willingly deceive ourselves — and not pathologically, but for fun! Pretending is self-illusion. When we pretend, we choose to misapprehend the world in order to enjoy an alternate reality. In fact, all games are illusion: in a schoolyard game of tag, what does it mean to be "it" if being "it" is not an illusion? If the participant who is "it"

[1] Understanding illusion as play is borne out by its etymology. Illusion comes from the Latin *lūdere*, meaning to play, to amuse, to mock, to mimic, to tease, to deceive, and to trick, among others.

no longer wants to be "it," they can simply cease to play — cease to pretend, cease to weave self-illusion — and that alternate reality disappears.

Yet people sometimes use illusions for nefarious purposes. Like the magician who produces a coin from behind your ear, some will produce illusory coins and say they are real to manufacture money where there was none before. These coins look and feel precisely the same as their authentic counterparts, but they have no legal value. Counterfeit documents mimic the signs and seals of real documents in order to produce an illusion of authority. Counterfeit paintings, down to their very brushstrokes and paints, pass themselves off as the real deal though they are mere illusions. Similarly, cleverly manipulated images construct an illusory reality in which events that did not happen in the real world happened.

The grammar of illusions highlights their internal mechanism. Illusions speak for themselves. The counterfeit coins do the work of maintaining the illusion, not the counterfeiter. By the time that a shopkeepers asks if certain coins are counterfeit, the illusion has already failed. Likewise, a counterfeit painting mimics authenticity down to its very materials because the word of its seller has little bearing when it comes to maintaining the illusion. Instead, illusions require no outside explanation. They contain their own evidence. Realistic computer imagery convinces us it is real because we are seeing it with our own eyes, and a magician's tricks would not mislead us in the slightest if we could not perceive them. In other words, illusions do not merely describe a world contrary to reality. They illustrate and perform it. As speech, they actually construct that world just like other performatives, such as saying "I do" in a wedding (Austin, 1975). In the words of countless literature teachers, illusions *show* us an alternate world instead of *telling* us of it.

It is here that deepfakes snugly fit in to the field of illusion. A deepfake is an illusion, not a lie. It tells a lie, but it is not one itself.[2] If a liar says something that is not true, they necessarily have no evidence for their claim *unless* they construct an illusion to support what they say. In that case, the liar's speech is secondary. It is the illusion that does the heavy lifting of popping a new, alternate reality into existence. A deepfake that purports to show President Obama cursing, making pop culture, references, and warning the world about deepfakes differs categorically from a statement that claims that happened. The deepfake is an illusion, whereas the statement is a lie. The deepfake counterfeits Barack Obama's speech, assuming his guise and taking his voice in order to present a claim. The lie, however, is just that: a lie. It does not convince us of its truth; the video of Obama saying those things does.

[2] Even though an illusion is not a lie itself, it is not an illusion if it does not tell a lie or somehow evidence something that is not real. Hence, illusions are always somehow made in relation to the truth. A self-illusion that the floor is lava is not an illusion if the floor is *actually* lava — nor is the situation enjoyable.

Deepfakes threaten our society because of how potent they are as illusions. In a digital world, we have become accustomed to the movie screen's illusions — no movie-goer seriously thinks that *The Hobbit's* Smaug flies through New Zealand's majestic mountains and hoards gold deep within an ancient city of rock. Moreover, we have slowly become able, with middling efficacy, to pick out counterfeit still images without technological assistance. Sophisticated forgeries can still fool even the most experienced fact-checkers. However, we have never before been subject to realistic simulation of a human face or human voice, much less a simulation that looks and sounds exactly the same as someone else. When seeing and hearing are believing — and what more do we have to fall back on? — deepfaked, illusory video and audio strike at our very ability to find the truth.

Categorizing deepfakes as "counterfeit" speech is not a new idea (Green, 2019, p. 1452). In the context of political campaigns, which I discuss below in Section 2.B, malicious deepfakes pose a dangerous problem, and Rebecca Green terms such videos "counterfeit campaign speech." However, we cannot understand deepfakes as solely counterfeit speech since counterfeits are simply one kind of illusion. Much as impersonators and satirists weave illusions to make a joke or a political point, deepfakes present a similar, if more technologically advanced, opportunity. It is not hard to imagine using deepfake technology for all sorts of benign or even beneficial illusions, such an app that allows you to generate video of yourself speaking when provided with text[3] or a clever video conferencing utility that only transmits audio and recreates a realistic face on the other end to reduce bandwidth use. Though these examples barely scratch the surface of possible deepfake applications, they do illustrate that we cannot understand deepfakes only as counterfeits. In some cases, they are the digital equivalent of a counterfeit coin, but in others, they are analogous to a well-practiced magic trick or a useful tool. In other words, they are illusions.

As extraordinarily true-to-life illusions, deepfakes embody a post-truth politics just as low highway underpasses materialize class exclusion (Winner, 1980, p. 124). They are "designed and built in such a way" that they produce (or have the potential to produce, should they become widespread) a complete breakdown in our ability to access the truth (Winner, 1980, p. 125). Considering how much of our lives now occur on the internet, we need some sort of ground truth to fall back on. Until now, we have always been able to trust videos and audio. In fact, even playful or helpful deepfakes present this threat to society. It does not matter who uses deepfakes nor how they use them when considering their epistemological effects. Instead, their destruction of the truth is built into

[3] This sort of app may be particularly helpful for people suffering from ALS and other conditions that impair speech. Instead of speaking in a computer-generated voice, patients could use technology to speak in their own voice long after they are able to produce understandable sounds.

their very form. Deepfakes are the epistemological equivalent of the nuclear bomb except this time everyone has the launch codes (Johnson & Diakopoulos, 2021).

Deepfakes in the wild

Pornography

As mentioned in the introduction, the first-noted and most widespread use of deepfakes is in pornography (Ajder et al., 2019, pp. 1–2). From Gal Gadot, a renowned celebrity, to ordinary women living non-descript lives, anyone is subject to potential deepfake porn if enough pictures of their face exist online. About a month after Vice's Samantha Cole discovered deepfakes on Reddit, her internet sleuthing turned up countless videos, made by a variety of users, that inserted celebrities' faces into pornography (Cole, 2018). Her articles' titles, both of which play off the double meaning of "fucked," clarify what is at stake in deepfake pornography.[4] When your face is swapped into a pornographic video, the results are just as if you had participated in that video. Your sexual agency is compromised as you are shown, absolutely realistically, engaging in any number of sex acts. It becomes nearly impossible to prove you did not participate, since video evidence of your participation exists right there. Even if you do manage to prove it, the video has already affected your reputation, your mental health, and your sexual agency. Deepfake pornography does not merely depict its targets sexually. Rather, it manufactures a convincing illusion of sex that is as real in its effects as actual sex. An illusion is real when it succeeds. Counterfeit money becomes real in every meaningful way when it is accepted in place of authentic coins. Illusory, deepfake pornography becomes real pornography when it acts on the world in the same way that real pornography would.

According to Danielle Keats Citron, deepfake sex videos "hijack people's sexual and intimate identities" and abridge their sexual privacy and agency (Citron, 2019, p. 1921). Though she writes that deepfake pornography is not the same as the nonconsensual distribution of explicit images (commonly known as "revenge porn"), she ascribes the same harms as revenge porn, sextortion, "up-skirt" photos, and nonconsensual recording to deepfakes. All of these acts invade their target's sexual privacy, traumatizing them in the process. Citron's examples include victims of deepfakes and other invasions of sexual privacy feeling unable to walk outside their homes without fearing that someone will recognize them from a pornographic video, suffering from recurrent feelings of exposure and vulnerability, and living in a state of "visceral fear" (Citron, 2019, pp. 1924–1926). Women whose faces appeared in pornographic videos made by ex-partners, unknown abusers, and random men in search

[4] The first meaning uses "fucked" to mean that one is unable to recover from a horrible thing, such as being "finished" or "done for." The second is the past participle of the verb "to fuck."

of a good target suffered severe, lasting emotional harm; psychological distress; reputational harm; and even job loss after those videos became public.

Deepfake pornography does not affect all people equally. Though the technology is effective against men and women alike, it is most commonly against women when it comes to pornography (Ajder et al., 2019, p. 2). As Citron points out, other acts that rip through one's sexual privacy, like nonconsensual pornography and cyber-stalking, are similarly gendered (Chesney & Citron, 2019, p. 1773). As such, sexual deepfakes fall squarely into Catharine MacKinnon's framework of sex-based group defamation-as-discrimination (MacKinnon, 1993, p. 99). In addition to the individual harms of deepfake pornography, illusory sex videos also constitute a gendered attack against women in general that manifests itself in our social reality. For instance, the general threat that anyone may make deepfake sex videos of anyone else constrains the actions of women who want to avoid those harms. Moreover, the power that men have to alienate the sexual privacy and agency of anyone they meet structures every interaction and permeates every space, including the home (Dodge & Johnstone, 2018). Deepfakes enable and sharpen intimate partner abuse, workplace sexual harassment, and other discrimination and hate.

MacKinnon's ideas also illustrate the nature of deepfaked sex videos themselves. A deepfake video is not a depiction of imagined nonconsensual sex, with imagined harms. It is a violent attack on one's sexual agency. It is as real as sexual assault and as impactful as it too. Just as rape survivors live through daily re-traumatization simply by existing in their bodies, victims of deepfake pornography feel exposed, vulnerable, and taken advantage of by very act of being seen by others. The existence of an illusion of sex becomes as powerful as sex itself.

Political misinformation and disinformation

In 2018, a few months after Samantha Cole broke the story about deepfake porn, former president Barack Obama warned the world about the possibilities of deepfakes turning our democracy into a "fucked-up dystopia" according to a video of his statement from BuzzFeed News (Sosa, 2018). However, Obama never said those words, at least not in a public address.

The end of the video reveals the illusion: Obama shrinks to one half of the screen and Jordan Peele, a well-known director, actor, and Obama impersonator, appears in the other half. The rest of the video subjects you to a surreal experience in which Jordan Peele's voice is Obama's and Obama's is Jordan Peele's. Both make the mouth movements required to produce the speech, and both make hand gestures that could reasonably accompany the audio.

While this deepfake used Obama's visage solely to communicate an educational point and drum up discussion, and clearly did not constitute misinformation, it captures the most obvious problems that deepfakes pose for democracy. If anyone can produce convincing video of politicians saying anything, then how is it possible to vote for our representatives? How is it possible to know what they do and do not believe? Nowadays, when a million-person political community ranks on the smaller side of the scale, digital communication is crucial to sustaining democracy. And deepfakes strike at the last remaining digital media — videos — free from convincing manipulation.

Much of the available literature on deepfakes has focused on their political implications. Rebecca Green, who I mentioned in the section on illusions, has dubbed deepfakes aimed at politicians "counterfeit campaign speech" since they fake "political candidates' identities, actions, words, and images" (2019, p. 1450). Key to her definition, which she later argues represents a class of speech that we should prohibit, is the mechanism of illusion. She distinguishes selectively edited videos from deepfakes and other counterfeit campaign speech because simply showing the full clip destroys an edited clip's deception (Green, 2019, p. 1452). Edited videos are deceptive, but not illusory. On the other hand, counterfeit campaign speech provides incontrovertible evidence of its fabricated claims. It is not merely a lie; it is a convincing alternate reality (Green, 2019, p. 1454).

Green identifies three types of harms of counterfeit campaign speech. First, she considers the harms that it does to voters. For Green, deepfakes deprive voters of their right to vote, which is actually a right to vote based on true knowledge of politicians' stances (Green, 2019, p. 1458). A voter who is uninformed or misinformed is unable to exercise their right to vote since they do not know what they are voting for. Deepfakes, by constructing an illusion that may guide voters in place of reality, and by subverting any attempt to destroy that illusion, prevent voters from exercising their agency and autonomy while voting.

Unlike lies, misleading statements, and bullshit,[5] all of which pop up frequently in hard-fought campaigns, deepfakes do not merely misinform voters. They make it difficult, if not impossible, to find the truth. Already, people reject corrections to classic political misinformation that confirms their prior views (Flynn et al., 2017, p. 130), and beliefs that people later recognize as false continue to affect their actions (Thorson, 2016). Beliefs based on false statements are potent while they are active and remain so after correction, but what happens when those beliefs are based on illusory realities instead of falsehoods? How easy will it be to convince someone that a politician did not actually say something that a deepfake shows them saying? Based on our experience with political misinformation and disinformation more broadly, it will border on the impossible.

In a perverse twist, deepfakes make it easier for candidates to discount real videos that depict them in compromising situations, like the infamous Access Hollywood tape of former President Donald Trump (Hasen, 2019, p. 543). When we cannot believe anything we see, politicians can easily

[5] Bullshit is a particular class of speech first theorized by Harry Frankfurt (2005). Bullshit is the speech of someone who does not care whether they speak the truth.

claim that real videos are actually fake. Dubbed the liar's dividend, the same dynamic motivates Trump's repeated fake news claims. Since many voters distrust the media, Trump is able to escape accountability for his actions by claiming that the reporters who broke the story are lying (Chesney & Citron, 2019, p. 1785). Widespread counterfeit evidence only makes it easier for liars to deny true events.

The second object of harm that Green identified is the electoral process more broadly. Deepfakes and malicious political illusions reduce voters' faith in democracy and self-governance (Green, 2019, p. 1460). If enough people become convinced that it is impossible to find the truth, the suppositions underlying democratic government fall apart. The effects of deepfakes on individual voters, taken to their limit, become the dissolution of democratic society itself. Deepfakes' inherent politics are anti-democratic.

Moreover, deepfakes attack the basic institutions of democracy through other avenues than the ballot box. For example, when a journalist encounters video or audio evidence of an important story, how should they treat it when there is a possibility it is a deepfake? Though news agencies fact-check everything they are given, they also rely heavily on the fact that until now, it has been almost impossible to produce convincing illusory audio and video. Even the possibility of a news agency succumbing to a deepfake and spreading its illusion might chill its enthusiasm to fulfill its fact-finding mission (Chesney & Citron, 2019, p. 1784).

Repressive regimes have long tried to use illusory images to discredit political opponents, though these illusions can backfire in an interconnected world (Farid, 2011; Venger, 2016). However, deepfakes are a more convincing illusion than single images, and it is already easy for a well-funded government propaganda agency to manufacture humiliating, ruinous deepfakes of individuals who pose political problems. Such video and audio may very well make it extraordinarily difficult for opposition movements to get off the ground.

Deepfakes' other harms to democracy include sowing division, undercutting public safety, constraining diplomacy, and attacking national security (Chesney & Citron, 2019, pp. 1780–1784). In any situation where someone saying something incendiary would cause harm, deepfakes have the potential to manufacture that harm. Examples of this sort of situation abound. In light of the mass protests against police brutality in the summer of 2020, Chesney and Citron's examples of a deepfaked police chief yelling racial slurs or a deepfaked community leader ordering violence against police officers are especially convincing.

Deepfakes also pose a problem for the arbiters of truth in our democracy: the courts. Besides their ability to convince individual judges of different realities, a problem that scholar Richard Hasen dubs "siloed justices," deepfakes present an epistemological and practical problem during lawsuits (Hasen, 2019, pp. 563–566). If any video at all may be faked, how is it possible for judges and juries to discern true evidence from falsified evidence? Though complex and invasive solutions that involve blockchains and mathematical verification may work in some cases, such as police bodycams, some evidence will always be unverifiable (“Decoding Deepfakes,” 2020, p. 22). It remains to be seen if rules of evidence will incorporate deepfake-specific provisions for verifying the authenticity of video and audio evidence, but such provisions cannot exist if we cannot detect deepfakes.

Finally, Green considers deepfakes’ harms to candidates themselves. She argues that counterfeit campaign speech attacks candidates’ dignity and hijacks their identities in much the same way that deepfake pornography acts on its victims. This is not to say that that they suffer invasions of their sexual privacy, though this is indeed possible. A pornographic deepfake of a woman who is running for office would most likely cost her votes, harming her political campaign at the same time as it abridges her sexual privacy and agency.

In general, however, I claim that some of the harm of illusory speech involves the loss of identity itself, which also occurs in pornographic illusion. When a malicious actor produces a deepfake of a candidate, they subjugate the candidate’s identity for use in their illusion.[6] They deny that candidate control over their own speech. Speaking with the candidate’s stolen voice, deepfakers can spread false claims that tank their political careers and cause irreparable harm to their reputation. Though the law of defamation typically covers reputational harms from the point of view of the target — and these harms are real — voters also have an interest in using candidates’ reputations to guide their actions, which becomes impossible when deepfakes construct illusory reputations in place of real ones (Heymann, 2011, p. 1376).

Deepfakes’ harms to democracy, whether to voters, society, or candidates, only multiply when we consider the accessibility of deepfakes and modern speech. Deepfake technology itself is widespread and uncomplicated, and the social shortcuts provided by social media allow traditional mis- and disinformation to go viral. Those looking to affect politics or even make quick cash find that their fake news spreads like wildfire on social media (Hasen, 2018, pp. 206–208). Since social media platforms perform very little moderation, and these networks are open to everyone, including anonymous, fake accounts, speech becomes “cheap,” a descriptor first coined in a law review article by Eugene Volokh (1995). More than 20 years later, Richard Hasen revisited Volokh’s predictions and found a far darker world — our world — in which the pathologies of cheap speech were partway through tearing down American democracy. As deepfakes grow in popularity and use, they will act like nitrous oxide in the engine of viral fake news, intensifying its effects and speeding its dissemination.

[6] The theft involved is both physical and theoretical. Deepfakers literally steal their targets’ identities by using their images to train the GANs that generate deepfakes.

Deepfakes and the law of speech

In the face of the harms of pornographic, political, and porno-political deepfakes, few positive legal remedies exist in the United States: Congress held its first hearing in history on deepfakes in the summer of 2019 and has passed no laws to regulate them (Hao, 2019). Only three states have passed laws that introduce specific regulations for deepfakes. In 2019, Virginia amended its statute against nonconsensual pornography to include deepfakes; California instituted civil penalties for pornographic deepfakes and banned doctored media of politicians within 60 days of an election; and Texas criminalized creating and publishing a deepfake with the intent to injure a candidate or influence an election within 30 days of an election (Morris, 2019; Paul, 2019; "Virginia Bans 'deepfakes' and 'Deepnudes' Pornography," 2019).

Little judicial doctrine on deepfakes exists either. As far as I can tell, no one has ever filed suit over deepfakes or introduced true deepfakes as evidence in court.[7] In the absence of judicial or legislative direction, scholars who study the harms of deepfakes have proposed several possible remedies. Some only apply to pornographic deepfakes; others apply only to political deepfakes. However, since deepfakes are speech, the First Amendment may constrain our legal solutions.

The most relevant First Amendment precedent to the deepfake question is *New York Times v. Sullivan*, in which the Supreme Court held that the Constitution protects false speech about public officials from civil and criminal sanction unless the perpetrators spoke with "actual malice" (1964, p. 279). In 2012, a plurality of the Court further explained that falsity alone could not remove speech from constitutional protection in *United States v. Alvarez* (2012, p. 722). Based on the First Amendment doctrine of *NYT v. Sullivan* and *U.S. v. Alvarez*, a blanket ban on deepfakes would most likely not withstand judicial scrutiny.

Nor is such a ban even possible. Now that the technology exists online, it is impossible to bury it; it will forever exist on some corner of the internet regardless of the actions of the United States. Moreover, there is no certainty that such a ban is even desirable. Though deepfakes' very existence causes harm by undermining our connection to objective truth, our approach, now that they do exist, must aim at mitigation. Banning their use and development would simply move deepfake research behind closed doors, ruining our ability to develop effective deepfake detection. Counter-intuitively, now that deepfakes exist, our best chance for overcoming the epistemological problems they pose is to allow

[7] In the beginning of 2020, British news outlets reported that a woman had submitted deepfaked audio as evidence in a child custody suit (Swerling, 2020). However, the audio was only a "cheapfake," a term used to describe manipulated media produced by hand without the help of neural networks. Though cheapfakes still pose many of the same problems as deepfakes, they are not this essay's subject.

development and subsidize detection.[8] At the same time, we must not downplay deepfakes' capacity for harm. Deepfakes already cause actual damage to their targets and society. These harms cry out for regulation.

First, I propose, as others have before, that we strengthen and clarify civil remedies to defamation, privacy invasion, and other abuse rooted in deepfakes. Since deepfakes necessarily present falsehoods and often damage their target's reputation, they are governed by state defamation statutes ("Defamation," n.d.). Though these laws differ from state to state, *NYT v. Sullivan* constrains their application on First Amendment grounds, which makes it more difficult for public figures to successfully pursue these defamation cases. However, states should allow private individuals to sue for civil damages related to deepfake defamation, including pornography, that causes harm. Furthermore, states and the Supreme Court should allow public figures to sue for deepfake defamation without having to prove the "actual malice" standard.[9] Illusions that present false statements — especially sophisticated, hard to detect illusions — differ categorically from those false statements themselves (Blitz, 2018, p. 110). It is coherent to say that defamation of a public figure requires proving actual malice on the part of the perpetrator except when they manufacture evidence to support their claim; the counterfeiting involved turns the speech into fraud rather than simply falsehood (Green, 2019, p. 1483). Since satirical and educational deepfakes do not harm their audience through falsehood, they do not qualify as defamation and are not covered by these changes. Of course, it is not always easy to tell if something is satirical. Rather than attempting to understand the deepfake's intent, courts should award damages to deepfake targets if someone spreads harmful deepfakes without portraying them as satire. A deepfake created by a satirist, when stripped of its context and watermark, could very easily cause harm.

However, civil remedies cannot protect the most vulnerable from the harms of deepfakes. As Danielle Citron notes, civil cases require time, money, and most worryingly, identification. Hence, "many victims [of sexual privacy abuses] decline to bring civil suits because they do not want to expose themselves to their attackers any further" (Citron, 2019, p. 1930). Criminal sanctions, while complete with their own set of problems, can provide victims of deepfakes with protection even when they wish to remain anonymous. Particularly when it comes to nonconsensual deepfake pornography, we need criminal penalties to protect victims and "signal the significant harm that such invasions inflict" (Citron, 2019, p. 1931). If states and the federal government do not acknowledge the need for a

[8] Even though reality constrains us, the question of whether we should snap our fingers and eliminate deepfake technology if it were possible is an interesting one. Given the catastrophic societal pathologies that deepfakes may exacerbate in the next couple of years, I lean towards snapping.

[9] As a term of art, actual malice means "knowledge or reckless disregard for the possibility that [something was] false." (Chesney & Citron, 2019, pp. 1793–1794)

comprehensive sexual privacy statute, as Citron promotes, they should at least ban the creation and distribution of deepfake pornography (as well as other nonconsensual pornography). Catharine MacKinnon's observation that for such media, what it says is "exactly the measure of [its] harm" supports such a ban (MacKinnon, 1993, p. 22); there is no protection in the First Amendment for invading another's sexual privacy and using their image as sex, regardless of the fact that such an act does speak.

Criminal sanctions for deepfakes of candidates or government officials are equally prudent. Like civil penalties, we should exclude satirical, educational, and other non-damaging deepfakes from these statutes to avoid as many First Amendment issues as possible. Such statutes should focus on the harm done to American democracy and voters in their construction. In particular, a deepfake of a sitting government official has extraordinary potential for harm; just as law proscribes the impersonation of a government officer, deepfakes of them should be illegal as well (Citron, 2019, p. 1937).

However, the nature of social media hinders enforcing civil and criminal sanctions against the creators and distributors of deepfakes. Since so many accounts are anonymous, finding an actual perpetrator who can stand trial is not an easy task. Moreover, these websites are international, and a deepfake spread by a bot account somewhere outside the United States damages our democracy and our privacy just as much as a homegrown deepfake. In this gap, we must require internet platforms to act against deepfakes. According to Richard Hasen, "the government likely has the power under the Constitution to mandate a truth-in-labeling law requiring social media platforms [...] to label synthetic media" (2019, p. 549). The federal government should enact regulations requiring social media platforms to label viral deepfakes on their platforms. Though the automatic deepfake detection is not yet possible, human labelers can act as a stopgap measure until we develop better forensic tools.

Beyond labeling, Congress should require platforms to remove deepfakes that fit into the civil and criminal sanctions that I discussed above. Though Hasen does not go this far, such a requirement is necessary. Just because we cannot identify a perpetrator in the United States to prosecute for a given pornographic or political deepfake does not mean that the deepfake itself should freely spread around the internet. In particular, we should adopt Bobby Chesney and Danielle Citron's limiting amendment to Section 230, the federal law that governs internet platforms (2019, p. 1799). In its current form, the law has "evolved into a super-immunity" to legal liability for content that allows internet platforms to "ignore the propagation of damaging deep fakes "(Chesney & Citron, 2019, p. 1798). By introducing civil liabilities for platforms that continue to spread harmful deepfake videos, the federal government can encourage those platforms to take down illusory, dangerous content. In doing so, we can "minimize the most-serious harms that might follow from user-posted or user-distributed deep fakes" (Chesney & Citron, 2019, p. 1799). To ameliorate claims that such a provision would unduly burden internet platforms, we must be careful to not extend liability too far. With the current state of our deepfake detection technology, we cannot expect platforms to never host deepfakes, though such a choice may be prudent on their part. Instead, they should become liable when they fail to remove harmful deepfakes of nonconsenting individuals.

These four changes — allowing civil penalties for harmful deepfakes, criminalizing certain classes of deepfakes, enacting labeling regimes, and introducing civil liability for platforms that host deepfakes — will help reduce the harms that deepfakes inflict on us and our society. The federal government should also pre-emptively increase investment in deepfake detection to better protect us against illusory media that is convincing to the human eye. Finally, though, social norms must do what law cannot.

Now that deepfakes exist, no laws can prevent their further development. If it does not happen in an open-source codebase, it will happen deep in cyber-intelligence agencies. Rather than attempting to haul the water back up the waterfall by banning deepfake technology, we should build strong and effective digital literacy programs, discourage the use of deepfakes for anything not strictly necessary, and pour resources into detecting their use (Johnson & Diakopoulos, 2021). Like all illusions, some applications of deepfakes have high utility. However, given the epistemological threat they pose, it is unlikely that most deepfakes will hurt less than they help — even in the realms of satire, fun, and education. Our social norms should reflect that: it should not be common to make or distribute deepfakes. Where the law allows no easy remedy, our norms must help us protect our society from the harms that deepfakes inflict by their very existence.

Conclusion

Our worst fears about deepfakes — that they will lead to a complete breakdown of our access to objective truth — have yet to manifest. However, deepfakes' other harms are already present and real. Pornographic deepfakes strip away their targets' sexual privacy and sexual agency. Their targets (who are mostly women) often feel exposed, fearful that strangers might recognize them from a pornographic film they had no part in making and to which they never consented. Pornographic deepfakes, by using women for sex against their will, become nonconsensual sex itself, complete with the psychological trauma, reputational harms, and emotional pain usually associated with sexual abuse and rape.

Political deepfakes, though not necessarily distinct from pornographic deepfakes, also inflict reputational harm on their targets, stealing their identities to produce illusory self-defamation. More importantly, however, political deepfakes prevent voters from finding the truth about candidates and chill the journalism integral to a well-functioning democracy. Deepfakes of this nature will also force our courts to grapple with highly sophisticated illusory evidence, complicate the diplomatic process, and deepen partisan divides. Mis- and disinformation already poison our democracy without the help of advanced, hard-to-detect illusions. Deepfakes will only exacerbate these problems.

To stay on top of the problems posed by deepfakes, we should pass a battery of new laws targeted at their specific, actionable harms. From civil and criminal liability for harmful deepfakes to internet platform regulations, we have First Amendment-compliant legal remedies that will help those most hurt by deepfakes. We also need to discourage deepfaking more broadly, working through social norms in the areas where law cannot help us.

A comprehensive social response, with both norms and laws, is the only thing that can reduce the present and future harms of deepfakes. To quote Jordan Peele wearing Barack Obama's skin, we must take action if we are to escape our impending future: a deepfake-ridden, fucked up dystopia (Sosa, 2018).

References

Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). *The State of Deepfakes: Landscape, Threats, and Impact* (p. 27). Deeptrace. https://web.archive.org/web/20201004082457/https://regmedia.co.uk/2019/10/08/deepfake_report.pdf

Austin, J. L. (1975). *How to do things with words* (J. O. Urmson & M. Sbisá, Eds.; Second edition.). Harvard University Press.

Blitz, M. J. (2018). Lies, Line Drawing, and Deep Fake News. *Oklahoma Law Review*, *71*(1), 59–116.

Chesney, B., & Citron, D. K. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, *107*(6), 1753–1820.

Choi, C. Q. (2017, July 12). *AI Creates Fake Obama*. IEEE Spectrum: Technology, Engineering, and Science News. https://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/ai-creates-fake-obama

Citron, D. K. (2019). Sexual Privacy. *Yale Law Journal*, *128*(7), 1870–1961.

Cole, S. (2017, December 11). AI-Assisted Fake Porn Is Here and We're All Fucked. *Vice*. https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn

Cole, S. (2018, January 24). We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now. *Vice*. https://www.vice.com/en_us/article/bjye8a/reddit-fake-porn-app-daisy-ridley

Decoding Deepfakes: How Do Lawyers Adapt When Seeing Isn't Always Believing? (2020, April). *Oregon State Bar Bulletin*, *80*(6), 7.

Defamation. (n.d.). In *Wex*. Legal Information Institute. Retrieved September 27, 2020, from https://www.law.cornell.edu/wex/defamation

Dodge, A., & Johnstone, E. (2018). *Using Fake Video Technology To Perpetrate Intimate Partner Abuse* (p. 8). Without My Consent.

Farid, H. (2011). *Photo Tampering Throughout History*. https://web.archive.org/web/20110606065418/http://www.cs.dartmouth.edu/farid/research/digitaltampering/

Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics. *Political Psychology*, *38*(S1), 127–150. https://doi.org/10.1111/pops.12394

Frankfurt, H. G. (2005). *On bullshit*. Princeton University Press.

Gholipour, B. (2017, May 2). New AI Tech Can Mimic Any Voice. *Scientific American*. https://www.scientificamerican.com/article/new-ai-tech-can-mimic-any-voice/

Green, R. (2019). Counterfeit Campaign Speech. *Hastings Law Journal*, *70*(6), 1445–1490.

Hao, K. (2019, June 12). Deepfakes have got Congress panicking. This is what it needs to do. *MIT Technology Review*. https://www.technologyreview.com/2019/06/12/134977/deepfakes-ai-congress-politics-election-facebook-social/

Hasen, R. L. (2018). Cheap Speech and What It Has Done (to American Democracy). *First Amendment Law Review*, *16*, 200–231.

Hasen, R. L. (2019). Deep Fakes, Bots, and Siloed Justices: American Election Law in a “Post-Truth” World. *Saint Louis University Law Journal*, *64*(4), 535–568.

Heymann, L. A. (2011). The Law of Reputation and the Interest of the Audience. *Boston College Law Review*, *52*(4), 1341–1440.

Johnson, D. G., & Diakopoulos, N. (2021). What to do about deepfakes. *Communications of the ACM*, *64*(3), 33–35. https://doi.org/10.1145/3447255

Korshunov, P., & Marcel, S. (2018). DeepFakes: A New Threat to Face Recognition? Assessment and Detection. *ArXiv:1812.08685 [Cs]*. http://arxiv.org/abs/1812.08685

Lewis, M. B., & Edmonds, A. J. (2003). Face Detection: Mapping Human Performance. *Perception*, *32*(8), 903–920. https://doi.org/10.1068/p5007

MacKinnon, C. A. (1993). *Only Words*. Harvard University Press.

Morris, A. (2019, October 9). Texas is first state to ban political “deepfake” videos. *San Antonio Express-News*. https://www.expressnews.com/news/local/politics/article/Texas-is-first-state-to-ban-political-14504294.php

New York Times Co. V. Sullivan, 376 U.S. 254 (Supreme Court of the United States 1964).

Pan, Z., Yu, W., Yi, X., Khan, A., Yuan, F., & Zheng, Y. (2019). Recent Progress on Generative Adversarial Networks (GANs): A Survey. *IEEE Access*, *7*, 36322–36333. https://doi.org/10.1109/ACCESS.2019.2905015

Paul, K. (2019, October 7). California makes ‘deepfake’ videos illegal, but law may be hard to enforce. *The Guardian*. https://www.theguardian.com/us-news/2019/oct/07/california-makes-deepfake-videos-illegal-but-law-may-be-hard-to-enforce

Sosa, J. (2018, April 17). *You Won’t Believe What Obama Says In This Video!* 😉. Monkeypaw Productions. https://www.youtube.com/watch?&v=cQ54GDm1eL0

Swerling, G. (2020, January 31). Doctored audio evidence used to damn father in custody battle. *The Telegraph*. https://www.telegraph.co.uk/news/2020/01/31/deepfake-audio-used-custody-battle-lawyer-reveals-doctored-evidence/

Thorson, E. (2016). Belief Echoes: The Persistent Effects of Corrected Misinformation. *Political Communication*, *33*(3), 460–480. https://doi.org/10.1080/10584609.2015.1102187

United States v. Alvarez, 567 U.S. 709 (Supreme Court of the United

States 2012). https://advance-lexis-com.ezproxy.amherst.edu/api/permalink/cb0d7390-e267-4725-8161-8f6eafb3e603/?context=1516831

Venger, O. (2016). When Shaming Backfires: The Doublespeak of Digitally-Manipulated Misogynistic Photographs. *CyberOrient*, *10*(1), 61–86.

Virginia bans “deepfakes” and “deepnudes” pornography. (2019, July 2). *BBC News*. https://www.bbc.com/news/technology-48839758

Volokh, E. (1995). Cheap Speech and What It Will Do. *The Yale Law Journal*, *104*(7), 1805–1850. JSTOR. https://doi.org/10.2307/797032

Winner, L. (1980). Do Artifacts Have Politics? *Daedalus*, *109*(1), 121–136.