VIRGINIA JOURNAL OF LAW & TECHNOLOGY

SPRING 2020 UNIVERSITY OF VIRGINIA VOL. 23, NO. 01

Deepfakes and the Weaponization of Disinformation

NINA I. BROWN†

† Assistant Professor, S.I. Newhouse School of Communications; J.D. Cornell Law School; B.S. Syracuse University. Many thanks to Kendra Peterson for her assistance in the preparation of this work.

ABSTRACT

Deepfakes are the latest weapon in the war against truth. Driven by advances in artificial intelligence, deepfake technology makes it remarkably easy to create realistic videos depicting events that never happened. Legislators, law enforcement officials, and private citizens have growing concerns about the potential for deepfakes to incite violence, target individuals, or disrupt elections.

These harms are not speculative. Deepfake nonconsensual pornography has already appeared online, and the United States has experienced foreign interference with our democratic discourse and elections. And the threat of further disruption is compounded by the unchecked power of social media.

In addition, the technology is evolving at such an alarming rate that it is increasingly difficult to distinguish deepfakes from authentic videos. Experts predict that a competent detection tool will not be available for years, if not decades. This has driven legislators throughout the U.S. to propose legislation that targets deepfakes, and their likely distribution channels—social platforms.

Deepfakes are frightening, but so is the rush to regulate them. Legislation requires careful deliberation, particularly when it targets at emerging technology. This is particularly true where, as here, there are positive uses for deepfakes, such as entertainment and parody, that come with strong First Amendment protections. Any solution needs to balance these factors and account for the fact that the technology—and its likely applications—will continue to evolve.

This article considers whether existing legal frameworks can effectively deter and punish those who create and distribute harmful deepfakes, or whether additional legislation is necessary.

TABLE OF CONTENTS

I. INTRODUCTION

When video footage of President Obama cursing and calling President Trump a derogatory name appeared online in 2018, it caused a stir. Surprisingly, the reaction to the video was not due to the content of the disparaging remarks, but to the fact that the video was fake.

Actor and writer Jordan Peele had created the fictitious video using artificial intelligence ("AI") software. Known as "deepfake" software, this type of AI makes it possible to create video and audio content of people taking actions and having conversations that never happened.[1] Peele made the video as a warning. He wanted to raise public awareness about deepfakes, and caution viewers not to believe everything they see online.

Discerning viewers of Peele's video could detect clues that suggested that it was inauthentic. In the video, President Obama's lip movement seems occasionally out of sync, or blurred.[2] Peele supplied the former President's voice audio, which is similar to President Obama's, but noticeably distinct from it. Because the video was created as a warning about misinformation, the tone recalls that of a classic public service announcement ("PSA"): the video opens to President Obama warning that society is "entering an era in which our enemies can make it look like anyone is saying anything at any point in time."

Since the video's launch, deepfake technology has continued to improve, removing many of the "tells" present in Peele's PSA. Newer deepfakes are crisp, clean and resemble

---

[1] Kaylee Fagan, *A viral video that appeared to show Obama calling Trump a 'dips—-' shows a disturbing new trend called 'deepfakes,'* BUS. INSIDER (Apr. 17, 2018), https://www.businessinsider.com/obama-deepfake-video-insulting-trump-2018-4.

[2] BuzzFeedVideo, Y*ou Won't Believe What Obama Says In This Video*, YOUTUBE (Apr. 17, 2018), https://youtu.be/cQ54GDm1eL0.

actual video footage. Most employ realistic voice duplication.[3] And they are easy to make. At the time of this writing, several deepfake programs are freely available to anyone with a computer. Although deepfakes are very much in their infancy, "this type of production carries immense potential to be indistinguishable from real-life videos."[4]

This is why the video caused such a stir: deepfake technology is rapidly becoming more realistic and accessible. Its potential for misuse is striking. In a world where we have long embraced the belief that "seeing is believing," this technology poses a great threat. This threat is particularly acute in light of the increasing ability of individuals to quickly disseminate messages to wide audiences through social media.

Compounding this threat is the fact that the majority of U.S. adults report that they have trouble identifying whether information they find online is trustworthy.[5] Even "digital natives," or those who became proficient in using computers and the Internet from an early age, have difficulty evaluating the veracity of information on their social media platforms.[6] Despite their concerns about the accuracy of news shared on social media, most Americans (68%) continue to consume

---

[3] James Vincent, *The AI-generated Joe Rogan fake has to be heard to be believed*, VERGE (May 17, 2019), https://www.theverge.com/2019/5/17/18629024/joe-rogan-ai-fake-voice-clone-deepfake-dessa.

[4] Douglas Harris, *Deepfakes: False Pornography Is Here and the Law Cannot Protect You*, 17 DUKE L. & TECH. REV. 99, 102 (2019).

[5] John B. Horrigan, *The spectrum of digital readiness for e-learning*, PEW RESEARCH CTR. (Sept. 20, 2016), https://www.pewinternet.org/2016/09/20/the-spectrum-of-digital-readiness-for-e-learning/.

[6] Joel Breakstone, et al., *Evaluating Information: The Cornerstone of Civic Online Reasoning*, STANFORD DIG. REPOSITORY (Nov. 22, 2016), http://purl.stanford.edu/fv751yt5934.

news in this way,[7] and share stories online despite not having read them.[8]

Convincing deepfakes could pose a serious threat to private individuals, companies, and governments both locally and internationally. Indeed, lawmakers and leaders of U.S. intelligence agencies have already identified deepfakes as a significant threat to global stability.[9] On the international level, "[a] well-timed and thoughtfully scripted deep fake or series of deep fakes could tip an election, spark violence in a city primed for civil unrest, bolster insurgent narratives about an enemy's supposed atrocities, or exacerbate political divisions in a society."[10] Deepfakes could also be used to exploit or target individuals: bad actors could, for example, seek to extort individuals by threatening the sale and distribution of nonconsensual deepfake pornography, or deliberately inflict emotional distress on them by releasing such material.

These harms are not speculative. Deepfake nonconsensual pornography has already appeared online,[11] and the U.S. has experienced foreign interference with our

---

[7] Katherine Eva Matsa & Elisa Shearer, *News Use Across Social Media Platforms 2018*, PEW RESEARCH CTR. (Sept. 10, 2018), https://www.journalism.org/2018/09/10/news-use-across-social-media-platforms-2018/.

[8] Caitlin Dewey, *6 in 10 of you will share this link without reading it, a new, depressing study says*, WASH. POST (June 16, 2016), https://www.washingtonpost.com/news/the-intersect/wp/2016/06/16/six-in-10-of-you-will-share-this-link-without-reading-it-according-to-a-new-and-depressing-study/?utm_term=.9d4760ca1820 ("According to a new study by computer scientists at Columbia University and the French National Institute, 59 percent of links shared on social media have never actually been clicked.").

[9] Alfred Ng, *Deepfakes, Disinformation Among Global Threats Cited At Senate Hearing*, CNET (Jan. 29, 2019, 11:35 AM PST), https://www.cnet.com/news/deepfakes-disinformation-among-global-threats-cited-at-senate-hearing/.

[10] Robert Chesney & Danielle Citron, *Disinformation on Steroids: The Threat of Deep Fakes*, COUNCIL ON FOREIGN RELATIONS (Oct. 16, 2018), https://www.cfr.org/report/deep-fake-disinformation-steroids.

[11] Harris, *supra* note 5, at 99, 100.

democratic discourse and elections.[12] With a presidential election looming, pressure is coming from citizens, commentators, and legislators to address this emerging problem. Complicating the issue is the fact that deepfakes can also be used for positive purposes with strong First Amendment protections, such as entertainment and commentary. To be successful, any solution must balance these disparate factors and account for the fact that the technology—and likely the way it is used—will continue to evolve

This paper explores the challenges posed by deepfakes, and considers possible solutions. Section II discusses the potential harms associated with deepfakes, given the increasing ease with which they can be made and distributed. Section III examines the special role that social media plays in the spread and influence of deepfakes. Section IV critically evaluates some of the technological and legal solutions that have been proposed, including the possibility for government regulation of social media. Finally, Section V offers a tentative roadmap for addressing the potential harms while balancing countervailing interests.

II. THE POWER AND DANGERS OF DEEPFAKES

A. The War on What is Real

Experts have widely predicted that deepfake technology will advance rapidly during coming years, and their predictions have helped to spur concerns about the potential for its abuse. Not surprisingly, when deepfakes first emerged online, most

---

[12] Julian E. Barnes, *Russia Could Unleash Fake Videos During Election, Schiff Says*, N.Y. TIMES (June 4, 2019), https://www.nytimes.com/2019/06/04/us/politics/russia-election-hacking.html; *see also* Max Boot & Max Bergmann, *Defending America From Foreign Election Interference*, COUNCIL ON FOREIGN RELATIONS (Mar. 6, 2019), https://www.cfr.org/report/defending-america-foreign-election-interference; Lily Rothman, *Fear of Foreign Intervention in U.S. Politics Goes Back to the Founding Fathers*, TIME (Dec. 17, 2016), https://time.com/4604464/foreign-interference-history/.

involved the creation and manipulation of pornographic video content. The faces of adult film stars were replaced with the faces of actresses and other women who did not consent to appear in the videos.[13] On its own, this application of deepfakes has the potential to cause significant harm. As scholars Danielle Citron and Bobby Chesney note, "[c]onscripting individuals (more often women) into fake porn undermines their agency, reduces them to sexual objects, engenders feeling of embarrassment and shame, and inflicts reputational harm that can devastate careers (especially for everyday people)."[14]

There is no evidence that deepfakes have yet been deployed for unlawful uses other than nonconsensual pornography.[15] But this fact likely owes to the relative newness of the supporting technology. As deepfake technology matures and improves, it can potentially be abused in myriad ways. For example, a deepfake could be made to depict a politician engaging in an illicit act, and then posted online shortly before an election. A deepfake could be used to provoke mass panic, by depicting the President informing citizens of an imminent or ongoing attack on the U.S. A deepfake could be used to discredit the Supreme Court, by depicting one of the Justices admitting to having taken bribes. The list goes on.

Any one of these potential deepfakes could cause tremendous harm to their targets and to the public. But challenging the authenticity of deepfakes could actually cause even greater societal harm, by playing into a broader war on

---

[13] Megan Farokhmanesh, *Deepfakes Are Disappearing from Parts of the Web, But They're Not Going Away*, VERGE (Feb. 9, 2018, 9:00 AM EST), https://www.theverge.com/2018/2/9/16986602/deepfakes-banned-reddit-ai-faceswap-porn.

[14] Robert Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CAL. L. R. 1753 (2019).

[15] It has been alleged, though unconfirmed, that deepfakes played a role in a political crisis in Gabon in December 2018. *See* Ali Breland, *The Bizarre and Terrifying Case of the "Deepfake" Video that Helped Bring an African Nation to the Brink*, MOTHER JONES (Mar. 15, 2019), https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/.

reality. In each of the listed hypotheticals, the subject of the video would likely immediately report that it was fake, and assert that the events portrayed therein never occurred. But in many cases, this "retraction" would be insufficient to counter the resultant public misinformation. Some viewers of the deepfakes may take action in response to false information, before a retraction can be issued; others might reject evidence showing that the footage is fake, particularly if its content confirms their prejudices or preexisting assumptions.[16] This is unsurprising, given our innate tendencies to seek and interpret information in a way that confirms our preconceptions. These confirmation biases make us "psychologically primed to accept without question new information that confirms [such beliefs]."[17] In other words, a retraction won't stop people from believing and spreading the fake video.

Fake news stories circulated prior to the 2016 election serve as compelling examples, such as the lie that President Obama was not born in the U.S. Despite the complete lack of evidence to support this assertion, and the availability of substantial evidence contradicting it, 42% of Republicans surveyed in 2017 believed that President Obama was not born in the United States.[18] It stands to reason that a compromising deepfake of President Obama — confiding to an aide that he was indeed born abroad, for example — could have a more potent impact.

Therefore, the deepfakes most difficult to debunk will be those that "confirm" the substance of preexisting rumors or

---

[16] *See* Donie O'Sullivan, *When seeing is no longer believing, Inside the Pentagon's race against deepfake videos*, CNN BUS. (Jan. 2019), https://www.cnn.com/interactive/2019/01/business/pentagons-race-against-deepfakes/.

[17] Ari Ezra Waldman, *The Marketplace of Fake News*, 20 U. PA. J. CONST. L. 845, 851 (2018).

[18] Eric Zorn, *Polls reveal sobering extent of nation's fact crisis*, CHI. TRIB. (Jan. 5, 2017), https://www.chicagotribune.com/news/opinion/zorn/ct-polling-ignorance-facts-trump-zorn-perspec-0106-md-20170105-column.html.

conspiracy theories—that a politician is corrupt or abusive, for example. For the target of the deepfake to successfully prove a negative—that the depicted events never occurred—in the face of video evidence to the contrary will be exceptionally challenging, particularly among groups predisposed to distrust or dislike the target. There will be no shortage of groups susceptible to politically motivated deepfakes. This will have significant consequences. Elections may be tipped, unrest may ensue, and individuals may suffer personal losses. These are all very real concerns, but the greatest threat emerges as citizens begin to understand that the technology allows realistic video to be created and used for unlawful purposes. The public trust may be eroded when we cannot believe what we see.

This erosion will likely manifest in two distinct ways. First, it will empower people to deny actual events captured on video. "Deep fakes will allow individuals to live in their own subjective realities, where beliefs can be supported by manufactured 'facts.'"[19] To some degree, this phenomenon predates the spread of deepfakes. "Some people already question the facts surrounding events that unquestionably happened, like the Holocaust, the moon landing and 9/11, despite video proof [of their occurrence]."[20] Deepfake videos could be used to bolster the erroneous beliefs of contrarians and conspiracy theorists, further emboldening them and their followers.

Second, citizens in search of the truth may be unable to discern whether video evidence is reliable. Experts warn that "the doubt sown by a single convincing deepfake could alter our trust in audio and video for good."[21] If viewers cannot distinguish authentic videos from fabricated ones on their own, they will be disinclined to trust *any* video evidence, whether offered as part of a news story, or as evidence in a courtroom.

---

[19] Chesney & Citron, *supra* note 15 ("Particularly where strong narratives of distrust already exist, provocative deep fakes will find a primed audience.").
[20] O'Sullivan, *supra* note 17.
[21] *Id*.

It is worth noting that journalists will encounter the same predicament. As scholars Danielle Citron and Bobby Chesney write:

> As the capacity to produce deep fakes spreads, journalists increasingly will encounter a dilemma: when someone provides video or audio evidence of a newsworthy event, can its authenticity be trusted? That is not a novel question, but it will be harder to answer as deep fakes proliferate. News organizations may be chilled from rapidly reporting real, disturbing events for fear that the evidence of them will turn out to be fake.[22]

Even the most ardent supporters of journalism will be forced to question the authenticity of video and audio footage relied on by journalists, without additional evidence that the depicted events occurred. When people cannot tell the difference between what is true and false, it reduces trust in traditional media, "making it difficult for true stories to have impact."[23] "Put simply: a skeptical public will be primed to doubt the authenticity of real audio and video evidence. This skepticism can be invoked just as well against authentic as against adulterated content."[24] Imagine how different the impact would have been if the Access Hollywood audio recording where Donald Trump made lewd statements about women was released in the era of convincing deepfakes. Supporters of then-candidate Donald Trump would reject the recording as fake, his opponents would hold it out as true, and those unsure would be caught in the space where they don't know what to believe.

Our democracy depends on our ability to engage in intellectual discussion and debate that is founded on a shared

---

[22] Chesney & Citron, *supra* note 15.
[23] Waldman, *supra* note 18.
[24] Chesney & Citron, *supra* note 20.

set of truths.[25] But in the U.S., truth has been under attack in from those who disagree with it. "[T]he simple introduction of empirical evidence can alienate those who have come to view statistics as elitist."[26] Even the President attempts to discredit unflattering news coverage by referring to it as "fake."[27] The introduction of realistic deepfakes will further fracture what little remains of these shared truths.

III. THE GATEKEEPERS WITHOUT A GATE

A. The Potential for Quick Distribution is as Concerning as Deepfakes Themselves

When the first pornographic deepfakes began to appear online, many platforms where they were posted and shared (or likely to be) responded by banning pornography containing face-swapping, one of the key markers of a deepfake.[28] This was an important step in curbing the potential for the deepfakes to spread. But it was not enough. Even if the ban was

---

[25] Robert C. Post, *Data Privacy and Dignitary Privacy: Google Spain, the Right to Be Forgotten, and the Construction of the Public Sphere*, 67 DUKE L.J. 981, 1005 (2018).
[26] Chesney & Citron, *supra* note 15.
[27] *See* Priscilla Alvarez, *CNN Takes on Donald Trump's 'Fake News' Label*, ATLANTIC (May 2, 2017), https://www.theatlantic.com/politics/archive/2017/05/cnn-trump-feud/525096/; Danielle Kurtzleben, *with 'Fake News,' Trump Moves from Alternative Facts to Alternative Language*, NPR (Feb. 17, 2017), http://www.npr.org/2017/02/17/515630467/with-fake-news-trump-moves-from-alternative-facts-to-alternative-language.
[28] *See* James Vincent, *Twitter is Removing Face-Swapped AI Porn from its Platform, too*, VERGE (Feb. 7, 2018), https://www.theverge.com/2018/2/7/16984360/twitter-ban-fake-porn-ai-face-swap; Adi Robertson, *Pornhub is the latest platform to ban AI-generated 'deepfakes' porn*, VERGE (Feb. 6, 2018), https://www.theverge.com/2018/2/6/16980920/pornhub-bans-deepfakes-fake-ai-celebrity-porn-video; Alex Hern, *Reddit bans 'deepfakes' face-swap porn community*, GUARDIAN (Feb. 8, 2018), https://www.theguardian.com/technology/2018/feb/08/reddit-bans-deepfakes-face-swap-porn-community.

successful in blocking all fake non-consensual pornographic content, those non-consenting individuals whose faces appeared in the videos still suffered harm the moment the video was created. The mere existence of a video that depicts them engaging in acts they never engaged in, without their consent, is harmful, even when it is not distributed. The psychological damage of knowing that one's face has been manipulated into a deepfake pornographic film would be profound. Thwarting the spread of deepfakes is important and helps to mitigate this harm, but does not eliminate it entirely.

Other potential abuses at the individual level, such as blackmail or identity theft may cause harm simply because the deepfake is created, or because it is targeted towards a small group of viewers on whom it will have a significant impact. The victims may doubt their ability to prove that the video is false, and make decisions to comply with the extortion based on their perceived inability to correct the misinformation. In this sense, deepfakes targeted at individuals are harmful not just because of the potential for wide dissemination, but because of the way they can manipulate the victim. Even if online sharing platforms banned deepfakes, it would offer little protection against this harm. This is not to suggest that such bans are unimportant—on the contrary, they are necessary tools in protecting individuals and fighting the spread of disinformation—but they alone cannot eliminate the harm, or even meaningfully reduce it.[29] The threat of widespread dissemination only compounds this harm.

This is not true of deepfake abuses that target groups of people. Deepfakes designed to disrupt elections or threaten public safety,[30] for example, would necessarily rely on wide

[29] Indeed, because the technology is publicly available, completely eradicating the possibility of these harms is unlikely. As will be discussed in Section C, *infra*, legal ramifications might reduce the number of abusive deepfakes targeted at individuals, but will not eliminate it altogether.

[30] One example of a deepfake that could threaten public safety might include false news videos seemingly from a reputable source that announce a nuclear attack.

distribution in order to have their desired impact. A deepfake of a politician engaging in an illicit act before an election, generated to sway voters in favor of their opponent, for example, will have limited impact unless it is spread to a significant part of the voting public. The easiest way to accomplish this broad dissemination would be through social media platforms like Facebook, Twitter, Reddit, and YouTube. Their services are free, accessible, and allow individuals to disseminate unfiltered information to a broad audience in real time. Even though most social media platforms' Terms of Use would likely prohibit the sharing of such deepfakes, in practice they would not prevent the initial distribution of the content itself.[31] In other words, although the content may eventually be removed for violating the Terms of Use, there will be a time lag between initial upload and removal. Removal can only occur after the content has been flagged (most often by a concerned viewer) and processed through an internal review—by which point, the video may have already spread sufficiently to have the intended impact.[32]

Consider the video of Speaker of the House Nancy Pelosi created and shared online in May 2019. A "Trump superfan" digitally manipulated a video of Speaker Pelosi to make it appear that she was drunkenly slurring her words, and posted it on Facebook.[33] Although the video was quickly identified as fake, it had already been tweeted by President Trump and viewed millions of times.[34] The Pelosi video was

---

[31] As will be discussed in Section IIIB, *infra*, social media companies have had a variety of responses to the circulation of false information on their platforms which in some cases may have led to its reduction, but not elimination.

[32] George Harrison, *Inkstagram, Painful, ugly and there FOREVER: your negative online posts are like a bad tattoo, says tech expert*, SUN (Nov. 22, 2017), https://www.thesun.co.uk/tech/4972145/social-media-tattoo-safety-online/.

[33] Simon Parkin, *The rise of the deepfake and the threat to democracy*, GUARDIAN (June 22 2019), https://www.theguardian.com/technology/ng-interactive/2019/jun/22/the-rise-of-the-deepfake-and-the-threat-to-democracy.

[34] *Id*.

not a deepfake: it was instead dubbed a "cheapfake," or "shallow fake" because it was distorted with a simple editing technique. But the incident highlighted just how fast fake videos, including deepfakes can spread, and the near-total discretion that social platforms exercise in deciding whether and how to respond.

This is one of the principle reasons that deepfakes have caused such alarm. The problem is not simply that the technology exists and is consistently improving, but rather that once created, deepfakes are easy to disseminate and difficult to eradicate. This "looming era of deep fakes will be different [than previous false video and audio content] because the capacity to create hyper-realistic, difficult-to-debunk fake video and audio content will spread far and wide."[35]

B. On Social Media, Seeing is Often Believing

Before the Internet democratized the spread of information, people relied heavily on traditional forms of broadcast and print media for news and information, and these typically implemented fact-checking or verification procedures. As the Internet has revolutionized the way we receive and share information, it "has also facilitated the spread of *mis*information because it obviates the use of conventional 'gate-keeping' mechanisms, such as professional editors."[36] There are no authentication filters within social media platforms that validate information before it is shared.[37] No step in the sharing process contemplates fact-checking or verification. When users consume information from non-traditional news sources, such as social media sites, the onus to perform this function shifts to them.

---

[35] Chesney & Citron, *supra* note 15.

[36] Stephan Lewandowsky, et al., *Misinformation and Its Correction: Continued Influence and Successful Debiasing*, 13 PSYCHOLOGICAL SCI. IN PUB. INTEREST 106, 110 (2012), http://www.jstor.org/stable/23484653.

[37] Although in recent months some social media companies have deployed tools to assist with the detection of fake news, these tools are buried in the interface in such a way that they are easy to miss.

Sometimes this task is relatively straightforward: for example, when users see that a story has been shared from a well-regarded and historically trustworthy news source like The Associated Press or ABC News, they may consume the information with little reason for skepticism about its accuracy. Other times, when the source is less reliable, verifying its accuracy is a greater hurdle for viewers, and in consequence, they often they skip the step entirely. This is true even when the information on its face appears dubious. As one commentator has written, "[s]topping to drill down and determine the true source of a foul-smelling story can be tricky, even for the motivated skeptic, and mentally it's hard work. Ideological leanings and viewing choices are conscious, downstream factors that come into play only after automatic cognitive biases have already had their way, abetted by the algorithms and social nature of digital interactions."[38]

People connected on social platforms tend to share similar sets of beliefs,[39] so social media users often see the same information repeatedly, as it is shared by different members of their social groups. This reinforces the trustworthiness of the information in two ways. First, people are predisposed to accept information as true when it is consistent with their preexisting beliefs.[40] Second, people are more likely to accept information as true when they are exposed to it repeatedly.[41] Users are more likely to accept false information as true upon repeated exposure to it—even when they initially rejected the information as false.[42] As one

---

[38] Benedict Carey, *How Fiction Becomes Fact on Social Media*, N.Y. TIMES (Oct. 20, 2017), https://www.nytimes.com/2017/10/20/health/social-media-fake-news.html.
[39] *See* Chesney & Citron, *supra* note 15.
[40] Lewandowsky, et al., *supra* note 37 at 112.
[41] *Id* at 113.
[42] David Z. Hambrick & Madeline Marquardt, *Cognitive Ability and Vulnerability to Fake News*, SCI. AM. (Feb. 6, 2018) https://www.scientificamerican.com/article/cognitive-ability-and-vulnerability-to-fake-news/; Lynn Hasher & David Goldstein, *Frequency and the Conference of Referential Validity*, 16 J. OF VERBAL LEARNING AND VERBAL BEHAV. 107 (1977).

journalist commented, regarding the conspiracy theory about President Obama's birthplace: "Over time, for many people, it is that false initial connection that stays the strongest, not the retractions or corrections: 'Was Obama a Muslim? I seem to remember that . . .'"[43] A deepfake video can cause people to form lasting opinions or beliefs, or reinforce them, particularly when viewed multiple times. As explained by journalist Kevin Roose:

> Online misinformation, no matter how sleekly produced, spreads through a familiar process once it enters our social distribution channels. The hoax gets 50,000 shares, and the debunking an hour later gets 200. The carnival barker gets an algorithmic boost on services like Facebook and YouTube, while the expert screams into the void. There's no reason to believe that deepfake videos will operate any differently. People will share them when they're ideologically convenient and dismiss them when they're not.[44]

The risk, of course, is that social platforms enable the instant delivery of content, without content verification.[45] Once shared, a video can do tremendous damage, even if it is subsequently proven to be a deepfake. The structure of the social web presents challenges to confronting the problem of deepfakes.

---

[43] Carey, *supra* note 39.
[44] Kevin Roose, *Here Come the Fake Videos, Too*, N.Y. TIMES (Mar. 4, 2018), https://www.nytimes.com/2018/03/04/technology/fake-videos-deepfakes.html.
[45] Ashley C. Nicolas, *Taming the Trolls: The Need for an International Legal Framework to Regulate State Use of Disinformation on Social Media*, 107 GEO. L.J. ONLINE 36, 42 (2018).

C. Engagement is Everything

Social platforms are profit-oriented organizations. Profits increase when users log in and view, click, read, share, and otherwise engage with their platforms.[46] To maximize user engagement, these platforms utilize algorithms to curate content that appeals to users' interests, appetites, and fears. These algorithms are fed user data (e.g., age, gender, location, occupation, etc.), data indicating how those users interact with content (for example, when they click, when they hover over content, whom they follow, etc.),[47] and other data which the platform has collected.[48] This algorithmic process is used to maximize user engagement, and deliver more of the type of content users already interact with.[49]

The algorithms are also fed data measuring how often a particular item is shared online, and this helps to determine what "gets circulated and what falls off the radar."[50] Collectively, the data collected and fed to these algorithms thus

---

[46] Katherine J. Wu, *Radical ideas spread through social media. Are the algorithms to blame?*, PBS SOCAL (Mar. 28, 2019), https://www.pbs.org/wgbh/nova/article/radical-ideas-social-media-algorithms; Sang Ah Kim, *Social Media Algorithms: Why You See What You See*, 2 GEO. L. TECH. REV. 147, 148 (2017) (explaining that maximizing user engagement leads to higher ad impressions and increased ad revenue).

[47] *Id.*; *see also* Andrew Griffin, *Facebook News Feed Algorithm to Track How Long Users Spend Reading Stories*, INDEPENDENT (June 15, 2015), http://www.independent.co.uk/lifestyle/gadgets-and-tech/news/facebook-news-feed-algorithm-to-track-how-long-users-spend-reading-stories-10320715.html [https://perma.cc/Q3L4-RWHS].

[48] Sang Ah Kim, *Social Media Algorithms: Why You See What You See*, 2 GEO. L. TECH. REV. 147, 148 (2017).

[49] Kalev Leetaru, *Is Social Media Curating Hate And Scouring The Web For Our Greatest Fears?*, FORBES (May 13, 2019, 1:54 PM), https://www.forbes.com/sites/kalevleetaru/2019/05/13/is-social-media-curating-hate-and-scouring-the-web-for-our-greatest-fears/#5c518d674cf3.

[50] Dewey, *supra* note 9.

determines what information users see and have the chance to interact with, and how that information is prioritized.[51]

Importantly, the more often a particular item of information is shared, the higher priority rating it receives. This is especially true where that item has been shared by a user's "friends," or others with strong connections within their social networks.[52] Thus, as "endorsements and shares accumulate, the chances for an algorithmic boost increase."[53] The result is that people in online communities, like "friend" groups, begin to see the same sets of information.

Unfortunately, this dynamic leads to the creation of echo chambers wherein "social media users are surrounded by information confirming their preexisting beliefs."[54] "Even without the influence of technology, people naturally tend to surround ourselves with information confirming our beliefs. Social media platforms supercharge this natural tendency by empowering users to endorse and re-share content."[55] To users, it appears that "the likes of the group inside the bubble represent the likes of the majority of people (because the group inside the bubble never sees anything contrary to its preferences)."[56]

It should not be surprising that the information in social feeds often serves to reinforce instead of challenge, users' preexisting views and ideologies—because the underlying algorithms are designed to do just that. And while this design

---

[51] *See*, *e.g.*, Chesney & Citron, *supra* note 11.
[52] Soroush Vosoughi, et al., *The Spread of True and False News Online*, 359 SCIENCE 1146, 1149 (Mar. 9, 2018), http://science.sciencemag.org/content/359/6380/1146/tab-pdf.
[53] Chesney & Citron, *supra* note 15.
[54] *See* Walter Quattrociocchi, Antonio Scala & Cass R. Sunstein, *Echo Chambers on Facebook*, JOHN M. OLIN CTR. L. ECON. & BUS. (June 13, 2016) (manuscript at 14), available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2795110.
[55] Chesney & Citron, *supra* note 15.
[56] Joanna Burkhardt, *Combating Fake News in the Digital Age*, 53 LIBR. TECH. REP. 8, 12 (2017).

feature may increase user engagement on a given platform, it helps to prevent social media from serving as a source of balanced and fair news coverage, and a forum for meaningful debate about public issues which might challenge users' preexisting beliefs.

D. When that News is Fake

Importantly, the algorithms that prioritize and filter content to maximize user engagement work exactly the same way when the shared content is false. The algorithms cannot detect those falsehoods or inaccuracies, so once misleading content begins to gain traction, it is often prioritized as something that users are interacting with, and distributed more frequently. Stories that are more likely to evoke a strong emotional response—positive or negative—are more likely to be shared.[57]

In this way, "[t]he information-sharing environment is well suited to the spread of falsehoods."[58] In a study about the spread of false news on Twitter over a ten-year period, researchers found that false information "diffused significantly farther, faster, deeper, and more broadly than the truth."[59] A study at MIT found that true stories took approximately six times as long as false stories to reach 1,500 people.[60] This is not to say that *all* false news spreads like wildfire on social media; but the potential for news stories to spread is greater when the stories are shocking but false, than when they are true. "Social media algorithms function at one level like evolutionary selection: Most lies and false rumors go nowhere, but the rare ones with appealing urban-myth "mutations" find psychological traction, then go viral."[61] Deepfakes are

[57] Lewandowsky, et al., *supra* note 37.
[58] *See*, *e.g.*, Chesney & Citron, *supra* note 11.
[59] Vosoughi, et al., *supra* note 53, at 1146-1151.
[60] *Id*.
[61] Carey, *supra* note 39.

frequently designed to be bold and shocking, and are therefore particularly likely to quickly go viral.

Once content with false information begins to spread on social media, the algorithms that govern the dissemination of that content actually begin to search for audiences receptive to it.[62] This distribution mechanism helps to explain why social media has played such a critical role in past disinformation campaigns. For example, in the lead-up to the 2016 U.S. presidential election, social media platforms served as incubators for false news stories.[63] Disinformation campaigns were so numerous and rampant that in the month preceding the election, more than 1 in 4 American adults (more than 65 million people) visited a fake news website,[64] and "the average American encountered between one and three stories from known publishers of fake news."[65] That fake news was (and remains) capable of spreading so easily on social platforms is in part because these networks are *built* for the rapid and wide dissemination of content generally, and shocking content in particular.

We are confronted with a perfect storm: the structure of social platforms, combined with the cognitive biases of human beings, creates fertile ground for the proliferation of deepfakes. [There is pressure from citizens, commentators, and lawmakers to confront this challenge.] Two forms of solutions—technological and legal—have been proposed. Sections IV-A and IV-C of this article more closely consider some of these proposed solutions.

---

[62] *Id*.
[63] Mike Wendling, *The (almost) complete history of "fake news,"* BBC (Jan. 22, 2018), https://www.bbc.com/news/blogs-trending-42724320.
[64] Andrew Guess, Brendan Nyhan & Jason Reifler, *Selective Exposure to Misinformation: Evidence from the Consumption of Fake News During the 2016 U.S. Presidential Campaign*, EUROPEAN RESEARCH COUNCIL (Jan. 9, 2018), *available at* https://www.dartmouth.edu/~nyhan/fake-news-2016.pdf (last visited Nov. 4, 2019).
[65] David M.J. Lazer et al., *The science of fake news: Addressing fake news requires a multidisciplinary effort,* 359 SCI. 1194, 1195 (2018).

IV. POSSIBLE SOLUTIONS

A. Technological Solutions

In both the public and private sectors, early efforts to develop tools to detect video manipulations are underway. Some of these efforts have been successful: a variety of detection mechanisms exist, and they are improving. But they still lag behind the sophistication of deepfakes, which continue to advance.

Researchers from a variety of institutions are working to develop detection algorithms. Harvard and MIT recently awarded $100,000 to the Rochester Institute of Technology (RIT) to create a deepfake-detecting software to help journalists identify fraudulent videos.[66] At the State University of New York at Albany, researchers have used algorithms that measure rates of eye-blinking to detect deepfakes. (Since the data used to create deepfakes comes predominantly from images of people with open eyes, those manipulated in deepfakes often blink at a lower than normal rate.[67]) Similarly, researchers at Purdue University are "using neural networks to detect the inconsistencies across the multiple frames in a video sequence that often result from face-swapping," and at [the UC system] a team has "developed methods to detect 'digital manipulations such as scaling, rotation or splicing,' that are commonly employed in deepfakes."[68]

---

[66] Victoria Hudgins, *Harvard and MIT Fund Deepfake Detection, Government Transparency AI Tools*, LEGALTECH NEWS (Mar. 19, 2019, 2:50 PM), https://www.law.com/legaltechnews/2019/03/19/harvard-and-mit-fund-deepfake-detection-government-transparency-ai-tools/?slreturn=20190414150849.

[67] Yuezun Li, Ming-Ching Chang & Siwei Lyu, *In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking*, UNIVERSITY AT ALBANY, SUNY (June 11, 2018), *available at* https://arxiv.org/pdf/1806.02877.pdf.

[68] John Villasenor, *Artificial intelligence, deepfakes, and the uncertain future of truth*, BROOKINGS TECHTANK (Feb. 14, 2019),

Researchers from Los Alamos National Laboratories are "creating a neurologically inspired system that searches for invisible tells that photos are AI-generated."[69] One of these "tells" is a discrepancy between the expected size of a video file, and its actual size: because deepfakes reuse visual elements from the dataset they are given, in many cases they contain less information than authentic video content.[70]

The Pentagon and its Defense Advanced Research Projects Agency ("DARPA") is working to develop systems to assess the integrity of video and audio content.[71] DARPA's aim is 'to make pivotal investments in breakthrough technologies for national security,"[72] and it has spent nearly $70 million on digital forensics technology to identify deepfakes.[73] DARPA has also teamed up with research institutions to confront this challenge.[74]

Inroads are being made in the private sector as well. Gyfcat, a gif-hosting platform, uses algorithms to "examine faces frame-by-frame to ensure nothing's been doctored" and

---

https://www.brookings.edu/blog/techtank/2019/02/14/artificial-intelligence-deepfakes-and-the-uncertain-future-of-truth/ (internal quotations omitted).

[69] Kaveh Waddel, *The impending war over deepfakes*, LOS ALAMOS NAT'L LAB. (July 22, 2018), https://www.lanl.gov/discover/features/top-media-stories/top-science-2018-22.php.

[70] *Id*.

[71] *Deep Intermodal Video Analytics (DIVA)*, INTELLIGENCE ADVANCED RESEARCH PROJECTS ACTIVITY, https://www.iarpa.gov/index.php/research-programs/diva (last visited May 7, 2018).

[72] *About DARPA*, DEF. ADVANCED RESEARCH PROJECTS AGENCY, https://www.darpa.mil/about-us/about-darpa (last visited July 9, 2019).

[73] Dan Robitzski, *DARPA Spent $68 Million on Technology to Spot Deepfakes*, FUTURISM (Nov. 19, 2018), https://futurism.com/darpa-68-million-technology-deepfakes.

[74] Matt Turek, *Media Forensics (MediFor)*, DEF. ADVANCED RESEARCH PROJECTS AGENCY, https://www.darpa.mil/program/media-forensics (last visited May 7, 2018); Donie O'Sullivan, *When seeing is no longer believing, Inside the Pentagon's race against deepfake videos*, CNN BUS. (Jan. 28, 2019), https://www.cnn.com/interactive/2019/01/business/pentagons-race-against-deepfakes/.

checks "whether a new gif has simply pasted a new face onto a previously uploaded clip."[75] These processes are relatively slow: removing content flagged by its algorithms can take as long as several days.[76]

According to Hany Farid, a computer scientist and digital forensics expert at Dartmouth College, the fight to detect deepfakes on a broad scale is a losing battle.[77] His rationale is simple. As researchers and developers make inroads in developing algorithms to detect deepfakes, the creators of deepfakes can work to evade the new verification mechanisms, and thereby make deepfakes harder to detect. As deepfake technology advances, detection will also become more challenging because the production value of deepfakes will likely improve, and so will innovations aimed at evading detection.[78] Even if researchers developed a system to accurately identify deep fakes *today*, that system would have to keep pace with the constant growth of the technology. Ultimately, "[t]he holy grail, a system that can automatically detect forgeries, is still well out of reach."[79] Farid warns that technology to comprehensively identify deepfakes is years away.[80]

Setting aside the real concern that this technology will never be able to "catch up" to the sophistication of deepfakes, as both technologies continue to evolve, there are obstacles to effective implementation. Even if a realistic deepfake is released to the public and identified as fake, it is unclear that the public will trust the detection software that flagged the

---

[75] Sarah Ashley O'Brien, *Deepfakes are coming. Is Big Tech Ready?*, CNN BUS. (Aug. 8, 2018, 11:16 AM), https://money.cnn.com/2018/08/08/technology/deepfakes-countermeasures-facebook-twitter-youtube/index.html.
[76] *Id*.
[77] Robitzski, *supra* note 74.
[78] Fillion, *supra* note 61 (noting that, as one researcher described, "the competition between generating and detecting fake videos is analogous to a chess game.").
[79] Waddel, *supra* note 70.
[80] *Id*.

video. The public may have little confidence in the party that supplied the detection software, and its claim that the video is a fabrication. A given individual's level of confidence will probably depend on a multitude of factors: their own ideological views and political biases, the original source of the deepfake and the forum through which they encountered it, the party announcing the video as a forgery, and more.

For example, imagine a deepfake video is released featuring the President engaged in a discussion with Russian agents about quid pro quos for interfering with the 2020 presidential election. Upon release of the video, DARPA detects the false content and flags the video as a deepfake. The Pentagon releases an official statement that the video, and the conversations contained therein, are fake. It is not clear which party the public would believe. Although some people will believe the Pentagon, many will not, and may instead conclude that the Pentagon—which is controlled by the Executive—is providing cover for the President. Many people are inclined towards conspiracy theories, and such theories already abound in respect of subjects ranging from the Holocaust, to the moon landing, to the terrorist attack of September 11th, 2001. Such people are often distrustful of government, and may be predisposed to find the content of such a video credible. An official statement from the Pentagon, alleging that the video "evidence" is fake, will likely fail to persuade such people. It may indeed serve to reinforce their belief that the video is legitimate, and induce them to conclude that the Pentagon is engaged in a cover-up.

In the deepfake arms race, there is more at stake than the threat of technological deception. It is the reality that deepfakes empower people to choose their truth.

B. A Second Technological Hurdle

If a technological mechanism for deepfake detection emerges sooner than experts expect, there still exists another significant hurdle: its implementation. To prevent deepfakes from spreading harms, it is necessary to detect and remove

them early on, in their forum or site of initial distribution. Harmful deepfakes will likely be shared in one of two ways: privately or publicly. Private sharing will include one-on-one communications used for the purposes of extortion. Importantly, the technological solution does nothing to eliminate the harms caused by private sharing, as this kind of distribution is direct, and there is no mediator to filter the media.

Many deepfakes will be publicly distributed, particularly those aimed at disrupting democratic discourse, undermining diplomacy, exacerbating social divisions, threatening public safety and national security, exacting revenge through the distribution of nonconsensual pornography, and otherwise inflicting widespread harms.[81] Technological solutions will prove useful only if they are implemented at the top of the distribution channel, whether that channel is traditional media outlets, or Internet platforms.

Traditional media outlets should, theoretically, be less likely to distribute deepfakes, because they adhere to journalistic standards of verification before publication. These processes of verifying the legitimacy of information, using more than one source, reduce the risk of spreading false information.[82] Even where journalists utilize social media for news gathering, they continue to place an emphasis on trusted sources and preexisting relationships,[83] and again verify information before publication, insulating their outlets from much of the risk associated with deepfakes.

---

[81] *See generally*, Chesney & Citron, *supra* note 15 (discussing harms caused by deepfakes).
[82] Ivor Shapiro, et al., *Verification as a Strategic Ritual*, 7 JOURNALISM PRACTICE 657 (2013).
[83] Nora Martin, *Information Verification in the Age of Digital Journalism*, UNIVERSITY OF TECHNOLOGY, SYDNEY (July 23, 2014), https://www.researchgate.net/profile/Nora_Martin2/publication/264121822_Information_Verification_in_the_Age_of_Digital_Journalism/links/53cf16310cf2fd75bc59b1a0.pdf.

The greater challenge is convincing companies that control online distribution channels to *voluntarily* employ the detection algorithms at the point of upload, and act to prevent the distribution of content flagged as problematic. The companies that control these channels are by and large social media platforms, where users upload and share content freely. For detection algorithms to have a serious impact, these companies would have to employ them as a filter in the upload process, to detect deepfake content before it can spread. And the use of filtering algorithms must be voluntary, because laws that *mandate* the detection and removal of false content on the basis of its falsity would likely run afoul of these companies' First Amendment rights.[84] Securing this voluntary compliance would likely prove to be a significant hurdle.

This is not to say that social platforms are unwilling or disinterested in stopping the spread of deepfakes; but this effort is associated with steep transaction costs that demand consideration. Even companies highly committed to stopping the spread of problematic deepfakes would face challenges to implementation. They would have to determine whether to ban *all* deepfakes, or just those that are abusive or have the potential to cause harm. And there are numerous benign and beneficial applications for deepfake technology. For example, YouTube may wish to allow reenactments of historical events created with this technology. Facebook might wish to allow parodies or non-harmful entertainment videos. Social media platforms may prefer a pro-speech policy that presumes that deepfakes are allowable *unless* their content would otherwise create a legal cause of action, such as fraud, defamation, etc.

Any policy short of a complete ban will require the platform to articulate a distinction between permissible and impermissible content. This is no simple task, particularly for platforms that already wrestle with determining which content

[84] Nina I. Brown & Jonathan Peters, *Say This, Not That: Government Regulation and Control of Social Media*, 68 SYRACUSE L. REV. 521 (2018).

ought to be blocked.[85] Twitter co-founder Biz Stone—who is no longer with the company—once commented that "[i]f you want to create a platform that allows for the freedom of expression for hundreds of millions of people around the world, you really have to take the good with the bad."[86] This meant that as Apple, Facebook, and Google deleted content posted by the far-right conspiracy site Infowars and its creator, Alex Jones, Twitter declined to ban Jones or Infowars, because they had not violated the company's lenient rules, despite spreading falsehoods that the Sandy Hook massacre was a hoax.[87] One month later, Twitter permanently suspended both "based on new reports of Tweets and videos posted [] that violate our abusive behavior policy, in addition to the accounts' past violations."[88] Twitter's decision followed intense political scrutiny and public outrage at its previous inaction, illustrating the pressure on these companies to draw a clear line of permissible conduct—and enforce it.[89]

Importantly, many social platforms have *not* drawn the line at publishing false information.[90] Facebook recently

---

[85] Cecilia Kan & Kate Conger, *Inside Twitter's Struggle Over What Gets Banned*, N.Y. TIMES (Aug. 10, 2018), https://www.nytimes.com/2018/08/10/technology/twitter-free-speech-infowars.html; Jessi Hempel, *Twitter's Latest Challenge: Deciding Who's a Terrorist*, WIRED (Jan. 8, 2016, 7:00 AM), http://www.wired.com/2016/01/twitters-latest-challenge-is-deciding-whos-aterrorist/ [http://perma.cc/HFX9-JRPZ] (noting that Twitter long "maintained one of the most liberal free speech policies among major social networks," and has struggled to draw a line at content to ban from its platform.).

[86] Michael Holmes, *ISIS Looking For Recruits Online*, WWLP (June 20, 2014, 11:00 PM), http://wwlp.com/2014/06/20/isis-looking-for-recruits-online/ [http://perma.cc/2E4Y-25PB].

[87] Kan & Conger, *supra* note 86.

[88] Avie Schneider, *Twitter Bans Alex Jones And InfoWars; Cites Abusive Behavior*, NPR (Sept. 6, 2018, 5:34 PM), https://www.npr.org/2018/09/06/645352618/twitter-bans-alex-jones-and-infowars-cites-abusive-behavior.

[89] *Id*.

[90] REDDIT, *Will you remove something defamatory about me or "my friend" from reddit?*, WWW.REDDIT.COM,

refused to remove the cheapfake video of House Speaker Nancy Pelosi, discussed *infra*, which was altered so that it appeared she was slurring her speech, because "false information alone does not violate the site's rules."[91] Within 24 hours the video had more than 2.5 million views.[92] In a statement defending its decision, Facebook said that it "believes 'reducing the distribution of inauthentic content' strikes the right balance between free speech and safety and concludes of certain misinformation."[93] YouTube came to the opposite decision, electing to remove the video for violating its Community Guidelines.[94] Balancing these competing interests (in the promotion of free expression, the promotion of truth, the prevention of harm and guarantee of public safety, and so on) represented a significant challenge for these platforms even before deepfakes were introduced.

Even if a social platform developed a policy that successfully balanced these interests, there would be no systematic way to distinguish between beneficial and potentially nefarious uses for deepfakes, because the underlying technology is the same for both.[95] To target just the problematic uses—while still allowing and encouraging positive ones—mandates an additional step in the review process, likely requiring human interpretation, that would be

---

https://www.reddit.com/wiki/faq#wiki_will_you_remove_something_defamatory_about_me_or_.22my_friend.22_from_reddit.3F (last visited July 9, 2019); *see also* Nancy Scola, *Facebook on fake Pelosi video: being 'false' isn't enough for removal*, POLITICO (May 24, 2019), https://www.politico.com/story/2019/05/24/facebook-fake-pelosi-video-1472413.

[91] Nancy Scola, *Facebook on fake Pelosi video: being 'false' isn't enough for removal*, POLITICO (May 24, 2019), https://www.politico.com/story/2019/05/24/facebook-fake-pelosi-video-1472413.

[92] *Id.*

[93] *Id.*

[94] *Id.*

[95] It would likely be simpler to target deepfake pornographic content. Algorithms are already utilized for detecting prohibited pornographic content; these could be combined with deepfake detection algorithms to ferret out just this type of content.

time-intensive and expensive. This is because drawing the line between "beneficial" and "harmful" uses will often be difficult, and require judgements that algorithms are incapable of making. (Arguably, unless a platform relied on attorneys to make such determinations, even moderators would frequently struggle to get this right.)

It may be unrealistic to expect that social platforms would voluntarily undertake such action, given that they collectively have billions[96] of users who upload significant content on a daily basis. On Facebook, for example, over 300 million pictures are posted each day.[97] On Twitter, users post over 500 million tweets per day.[98] 300 hours of video are uploaded every minute on YouTube.[99] This is by design: social media platforms were literally built to "sign up as many users as possible and have them posting, liking and commenting as often as possible."[100] They were not built to accommodate human content-filters or to moderate each post. Such endeavors may be impossible in light of the volume of information that is shared on these platforms.

---

[96] Dan Noyes, *The Top 20 Valuable Facebook Statistics – Updated July 2019*, ZEPHORIA (July 2019), https://zephoria.com/top-15-valuable-facebook-statistics/; Ben Gilbert, *YouTube Now Has Over 1.8 Billion Users Every Month, Within Spitting Distance of Facebook's 2 Billion*, BUS. INSIDER (May 4, 2018, 10:47 AM), https://www.businessinsider.com/youtube-user-statistics-2018-5.

[97] Dustin W. Stout, *Social Media Statistics 2019: Top Networks by the Numbers*, DUSTIN STOUT, https://dustinstout.com/social-media-statistics/ (last visited Nov. 2, 2019).

[98] *Twitter Usage Statistics*, INTERNET LIVE STATISTICS, http://www.internetlivestats.com/twitter-statistics/#trend [http://perma.cc/PJS7-DDPF] (last visited Nov. 2, 2019).

[99] Barbara Ortutay, *Social media and misinformation: It's a game of whack-a-mole*, ASSOCIATED PRESS (Dec. 18, 2018) https://apnews.com/0d02a1cec5b04638810372ba23e03ee3.

[100] David Seigel & Rob Reich, *It's Not Too Late for Social Media to Regulate Itself*, WIRED (Feb. 7, 2019), 9:00 AM), https://www.wired.com/story/its-not-too-late-for-social-media-to-regulate-itself/.

In consequence, there is a significant implementation problem. Were a viable technological solution available, social platforms would be forced to decide between using it to eliminate *all* deepfake content—even benign content—and employing it on a smaller scale.

At present, these concerns are academic: there are no detection algorithms available to address this threat. The ongoing efforts in the public and private sectors to combat the potential threat of deepfakes are critical, but a comprehensive technological solution is still years away. In light of this, scholars[101] and lawmakers[102] have called for legal solutions to address the issue. In an industry defined by rapid technological growth and change, these proposed solutions are associated with other challenges.

C. Legal Solutions

A natural response to the threat of deepfakes or any new technology that could be used in nefarious ways is to ban the technology altogether. However, such a strong reaction is unwarranted because deepfake technology is not inherently problematic. There are beneficial uses for the technology as well, and strong Constitutional protections for those positive uses complicate its regulation.

1. Beneficial Uses of Deepfakes Complicate Regulation

Like many technological advances—the Internet, drones, mobile phones—individuals and entities can utilize deepfakes to both positive and negative ends, and regulation

---

[101] *See* Robert Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CAL. L. REV. (forthcoming 2019).

[102] See, e.g., Malicious Deep Fake Prohibition Act of 2018, **S. 3805**, 115th Cong. (2018); Alfred Ng, *Deepfakes are a threat to national security, say lawmakers*, CNET (Sept. 13, 2018), https://www.cnet.com/news/deepfakes-are-a-threat-to-national-security-say-lawmakers/.

often becomes complicated. Beyond the myriad ways deepfake technology can be abused, beneficial uses abound across industries ranging from entertainment to healthcare. Individuals will be able to incorporate the technology to engage in satire and parody,[103] to critique the government and its leaders, or to insert themselves into favorite movies or news clips of historical events. Vocal avatars created by deepfake technology allow those who have lost their voice due to illness or injury the ability to continue using their own unique voice.[104] Hollywood has even embraced the technology, as it opens up the door to producing movies featuring stars that have died.[105] The technology would also allow for more realistic dubbing for foreign language films. And at least one adult film company sees a market for using deepfakes to include users in film scenes, particularly those "with physical limitations [who] could place themselves in sexual situations that would be impossible in real life."[106] Another use would be personalized advertising, "where the ads you see as you surf the web include you, your friends, and your family."[107] As deepfake technology still grows, so too will its positive applications.

---

[103] Indeed, deepfakes are already being used this way. *See* Helena Skinner, *French charity publishes deepfake of Trump saying 'AIDS is over,'* EURONEWS (Sept. 10, 2019), https://www.euronews.com/2019/10/09/french-charity-publishes-deepfake-of-trump-saying-aids-is-over; John Maher, *This was the year of the deepfake Nicolas Cage meme*, DAILY DOT (Dec. 27, 2018) https://www.dailydot.com/unclick/nicolas-cage-memes-deepfakes-2018/ (creating deepfakes where Nicolas Cage replaces actors in iconic films).

[104] Lyrebird AI is a project that partners with the ALS Association on Project Revoice to help people with ALS create a digital copy of their voice. *See Lyrebird AI*, DESCRIPT, https://www.descript.com/lyrebird-ai (last visited July 10, 2019).

[105] Of course, such uses trigger right of publicity concerns that will be discussed in Section C(3), *infra*.

[106] Jackie Snow, *An adult film company wants to put users into deepfake porn*, FAST CO. (Aug. 20, 2018), https://www.fastcompany.com/90221476/an-adult-film-company-is-putting-users-into-porn-with-a-deepfake-tool.

[107] Gaurav Oberoi, *Exploring DeepFakes,* MEDIUM (Mar. 5, 2018), https://goberoi.com/exploring-deepfakes-20c9947c22d9.

Banning deepfakes altogether would not only stifle these positive uses but would also raise insurmountable First Amendment hurdles. The First Amendment exists primarily to protect citizens against censorship of speech critical of the government.[108] While this type of speech is at the core of this freedom, its spread is naturally much broader. The First Amendment protects an individual's right to free speech not only to advance the discovery of truth and promote democratic self-government, but also to protect the individual's identity interest in self-expression.[109] Indeed, the Court has noted that "[s]uch expression is an integral part of the development of ideas and a sense of identity. To suppress expression is to reject the basic human desire for recognition and affront the individual's worth and dignity," and the First Amendment provides protection of "those precious personal rights by which we satisfy such basic yearnings of the human spirit."[110] Such a broad right to free expression includes parody and satire.[111] It is not limited to oral communication or writings, however, but also protects visual art such as films. The same rationale would

---

[108] Arizona Free Enter. Club's Freedom Club PAC v. Bennett, 564 U.S. 721, 754 (2011) (noting that the "whole point of the First Amendment is to protect speakers against unjustified government restrictions on speech."); McKee v. Cosby, 139 S. Ct. 675, 682 (2019) (*cert. denied*) (noting the "broad consensus" that the First Amendment protects "criticism of government and public officials.") (internal citations omitted).

[109] First Nat'l Bank of Boston v. Bellotti, 435 U.S. 765, 777 n.12 (1978) ("[T]he individual's interest in self-expression is a concern of the First Amendment separate from the concern for open and informed discussion."); Consol. Edison Co. v. Pub. Serv. Comm'n, 447 U.S. 530, 534 n.2 (1980) (observing that in addition to advancing the discovery of truth and promoting democratic self-government, "[f]reedom of speech also protects the individual's interest in self-expression"); Citizens United v. FEC, 130 S. Ct. 876, 972 (2010) (Stevens, J., concurring in part and dissenting in part) ("One fundamental concern of the First Amendment is to 'protec[t] the individual's interest in self-expression.'" (alteration in original) (quoting Consol. Edison, *supra*)).

[110] Procunier v. Martinez, 416 U.S. 396, 427 (1974), *overruled by* Thornburgh v. Abbott, 490 U.S. 401 (1989).

[111] Hustler Magazine, Inc. v. Falwell, 485 U.S. 46 (1988).

extend to deepfake recordings.[112] Thus, the creation of deepfakes itself is a protected First Amendment activity.

This is not to say that *all* deepfakes will be protected by the First Amendment. Those that contain defamatory content or cause emotional distress, for example, could be subject to liability. But the fact that a deepfake is a manufactured video and may contain falsehoods (i.e. President Obama *didn't really* make those remarks, Nicolas Cage *didn't really* star in those films, etc.) in and of itself does not weaken its First Amendment protection. Thus, any legislative ban on false or misleading deepfake videos would be Constitutionally problematic. The Constitution protects false speech, and the Supreme Court has been unwilling to carve out false speech as a category of speech undeserving of protection.[113] Even within some defined categories deemed to fall outside of the scope of the First Amendment, the Court has found value in protecting false statements when to do otherwise would have a "chilling effect."[114]

The Court's rationale is not that there is "constitutional value in false statements of fact" (indeed, it has explicitly rejected that idea),[115] but it recognizes that "[e]ven a false statement may be deemed to make a valuable contribution to

---

[112] Kaplan v. California, 413 U.S. 115, 119 (1973) ("As with pictures, films, paintings, drawings, and engravings, both oral utterance and the printed word have First Amendment protection."); *See, e.g.,* Kingsley Int'l Pictures Corp. v. Regents of Univ. of State of N.Y., 360 U.S. 684, 688 (1959) ("[T]he First Amendment's basic guarantee is of freedom to advocate ideas."); Superior Films, Inc. v. Dep't of Educ. of State of Ohio, Div. of Film Censorship, 346 U.S. 587, 589 (1954) (Douglas, J., concurring) ("Motion pictures are of course a different medium of expression than the public speech, the radio, the stage, the novel, or the magazine. But the First Amendment draws no distinction between the various methods of communicating ideas."); Joseph Burstyn v. Wilson, 343 U.S. 495, 502 ("[W]e conclude that expression by means of motion pictures is included within the free speech and free press guaranty of the First and Fourteenth Amendments.").

[113] United States v. Alvarez, 567 U.S. 709, 718 (2012).

[114] N.Y. Times Co. v. Sullivan, 376 U.S. 254, 299-301 (1964).

[115] Gertz v. Robert Welch, Inc., 418 U.S. 323, 340 (1974).

public debate, since it brings about 'the clearer perception and livelier impression of truth, produced by its collision with error.'"[116] A potential concern is that this marketplace of ideas theory in which falsehoods compete with truth until truth prevails may not function the same way on social platforms. Its algorithms spread deepfakes with more speed and breadth than posts depicting true events, thus making it difficult for the truth to compete.[117] The Court has recognized that in cases like this, when false factual statements cause or are likely to cause cognizable harm to other persons, some regulation might be appropriate.

The reality is that not all uses of deepfakes will be protected. Uses casting the subject-target in a false light or otherwise defaming them could intentionally or recklessly inflict emotional distress on those depicted. A benefit of having a legal framework in place is that it provides a dependable system of deterrence and restitution. If someone plans to engage in particular conduct, he or she can predict what the legal fallout will be and assess the risk accordingly. Likewise, those with an interest in preventing and redressing legal harms—the state and the individual victims—know what options are available to them once such harms have occurred. In addition, those who wish to use the technology for positive societal contributions are reassured that such usage is protected, encouraging them to utilize the technology instead of chilling their speech.

Even without regulation in place that specifically contemplates deepfakes, our laws already provide much of this framework—the emergence of deepfakes is not our first exposure to tools capable of manipulating our understanding of reality. Photo and video editing techniques have been around for decades and are well regulated under existing legal

---

[116] N.Y. Times Co. v. Sullivan, 376 U.S. at 279 n.19 (quoting John Stuart Mill, ON LIBERTY AND OTHER ESSAYS (World's Classics ed. 1991)).
[117] *See generally* Jared Schroeder, *Marketplace Theory in the Age of AI Communicators*, 17 FIRST AMEND. L. REV. 22, 30 (2018) (discussing how AI and tech communicators can flood the marketplace).

framework. Much of this will apply in the same way to deepfakes by providing a framework for people utilizing the technology to assess risk.

2. Criminal Laws

When deepfakes result in harm, there are a variety of laws that may apply to punish and provide restitution. For example, federal and state cyberstalking laws may apply when individuals are targeted and threatened or intimidated by deepfake content.[118] Humiliating and personally damaging deepfakes that are created as leverage to force individuals to engage in specific conduct would be subject to federal and state criminal extortion laws.[119] State harassment laws may also apply, though these vary widely regarding whether and to what extent specific provisions for online stalking or harassment are included."[120] Additionally, laws preventing fraud would apply broadly to deepfakes that were meant to deceive viewers and induce reliance.[121] Similarly, several states criminalize impersonation crimes.[122]

It may be the case that current laws do not adequately deter, punish, or provide restitution for the harms caused by deepfakes. If new legislation is necessary to fill these gaps, it will have to balance countervailing interests like First

---

[118] *See* 18 U.S.C.A. § 2261A (2018) (federal cyberstalking statute).
[119] 18 U.S.C.A. § 875(d) (2018) ("Whoever, with intent to extort from any person, firm, association, or corporation, any money or other thing of value, transmits in interstate or foreign commerce any communication *containing any threat to injure the* property or *reputation of the addressee* or of another or the reputation of a deceased person or any threat to accuse the addressee or any other person of a crime, shall be fined under this title or imprisoned not more than two years, or both.") (emphasis added).
[120] Emma Marshak, *Online Harassment: A Legislative Solution*, 54 HARV. J. ON LEGIS. 503, 514 (2017).
[121] *See, e.g.,* Eurycleia Partners, LP v. Seward & Kissel, LLP, 12 N.Y.3d 553, 559 (2009) ("The elements of a cause of action for fraud require a material misrepresentation of a fact, knowledge of its falsity, an intent to induce reliance, justifiable reliance by the plaintiff and damages.").
[122] *See* Chesney & Citron, *supra* note 102 (collecting statutes).

Amendment rights to survive Constitutional scrutiny. But even if there existed a well-conceived law that balanced these interests against the potential harms brought by deepfakes, there would likely still be significant hurdles to achieving justice. This is because the predictive nature of the law is powerful, but not perfect. For example, the existence of criminal and civil laws does not deter all crime. This is particularly true when it comes to computer crimes such as deepfakes[123] for two principal reasons.

First, those sophisticated enough to engage in online criminal activity often have the ability to remain anonymous.[124] This is borne out by current deepfake pioneers. The first person to publicly release deepfake code (a Reddit contributor named Deepfakes) operates anonymously.[125] So too does a second Reddit user who used Deepfakes' code to create FakeApp.[126] This cloak of anonymity dramatically reduces the ability of laws to regulate or deter deepfake abuses.[127] If perpetrators can avoid detection and, thus, sanctions, the law is of minimal consequence. To have a realized impact, the law would need an enforcement mechanism—some way to identify and target those responsible for the deepfake in question's creation or distribution.

---

[123] Brent Wible, *A Site Where Hackers Are Welcome: Using Hack-in Contests to Shape Preferences and Deter Computer Crime*, 112 YALE L.J. 1577, 1579-81 (2003) ("Scholars and policymakers have since proposed a number of deterrence strategies, from criminal sanctions to tort law and the architecture of the web itself, but none of these methods has proved successful at deterring criminal hacking.").

[124] Duncan B. Hollis, *An E-Sos for Cyberspace*, 52 HARV. INT'L L.J. 373, 378 (2011).

[125] Mark Wilson, *The War on What's Real*, FAST CO. (Mar. 6, 2018), https://www.fastcompany.com/90162494/the-war-on-whats-real.

[126] *Id.*

[127] *See*, *e.g.*, Hollis, *supra* note 125 at 374 (arguing that "[l]aw cannot regulate the authors of cyberthreats because anonymity is built into the very structure of the Internet. As a result, existing rules on cybercrime and cyberwar have little deterrent effect.").

The second reason criminal laws may have little deterrent effect on computer crimes is that the perpetrators can originate outside of the U.S. One commentator notes that the "fact that these [crimes] can originate in a country other than that of the victim(s) creates a jurisdictional barrier to accountability that allows perpetrators to further discount the chances of getting sanctioned."[128] Even the most punitive laws will have little effect on foreign actors outside of the reach of U.S. Courts. Disinformation campaigns deployed to disrupt elections and antagonize political divisions have targeted the U.S. from outside its borders. It is reasonable to assume that as deepfake technology improves, it will become an additional weapon employed in such attacks from abroad.

Legal solutions are also limited in that once the deepfake is created, the law cannot operate to stop its release or spread. It can only punish those who can be caught and brought within the reach of the law and perhaps deter others. This is not to suggest that laws against abuses of deepfakes are meaningless or unnecessary. They may indeed be important tools in deterring attacks and providing restitution to those injured. But they are far from a panacea to the threats posed by deepfakes, primarily because such laws are easy for sophisticated attackers to evade and thus may not bring about the desired deterrent or restitutive effect. We must keep in mind these limitations as we consider legal solutions to this problem.

3. Civil Liability

Currently, there is no civil law that provides redress for those specifically targeted by deepfakes. Instead, they must rely on a composite of civil remedies to be made whole. Defamation laws, for example, might compensate those whose likenesses have been used in ways harmful to their reputations. Likewise, in the states where it is available, false light would provide relief when videos cast subjects in an untrue manner

[128] *Id.* at 405.

that a reasonable person would find offensive.[129] Among the deepfakes most likely to spread virally are those that portray people behaving in shocking ways. When such people suffer emotional harm as a result of the deepfake, they may also find relief in emotional distress laws.[130] In cases where the subject of the deepfake is exploited to sell a product or service, Right of Publicity and commercialization laws may provide relief.[131] In addition, Section 43(a) of the Lanham Act would protect businesses against deepfakes that constituted unfair competition or included misleading advertising.[132]

That there is not a specific law targeting deepfakes will not prevent injured parties from seeking civil relief—they can make use of the many tort remedies listed above. Although calls for civil responses to deepfakes will no doubt occur, additional legislation is unlikely to improve the outcomes for victims of deepfakes. This is because there remain significant drawbacks to potential laws addressing these issues. For one, the injured party "will bear the responsibility to take the time,

---

[129] *See, e.g.*, Jackson v. Mayweather, 10 Cal. App. 5th 1240, 1256, 217 Cal. Rptr. 3d 234, 256 (Ct. App. 2017) (2017) (*cert. denied*) ("California courts have recognized four distinct types of right of privacy claims [including] false light . . .").

[130] *See* RESTATEMENT (SECOND) OF TORTS § 46 (AM. LAW INST. 1965).

[131] Such cases will likely be limited because "the harms associated with deep fakes do not typically generate direct financial gain for their creators." Chesney & Citron, *supra* note 102 at 35.

[132] Lanham Act § 43, 15 U.S.C. § 1125 (2012). Notably, although deepfakes often rely on images taken by third parties to create their works, copyright laws would also have limited ability, if any, to offer relief. Recall that deepfakes rely on data sets comprised of many individual data sources such as photos or videos to generate new content. Assuming arguendo that each photo or video was subject to copyright protection, it does not follow that the use of those images would be infringement. Deepfakes are *entirely new* creations based on those images—not the release of the images themselves. Although the fair use doctrine does not operate within bright line rules, a use of copyrighted works to create entirely new content would likely be considered a transformative use. Of course, finding that there has been a transformative use is not the end of the fair use inquiry. A court could find that the other factors of the test tipped so strongly against a finding of fair use that the use could be found to be an infringement. *See* 17 U.S.C. § 107 (1976).

energy and money to sue the deepfake creator, if that creator can actually be identified."[133] Identification of the creator may prove difficult or impossible if that party has taken sufficient steps to remain anonymous.

Even if the party responsible is identified, any litigation will expose the plaintiff to reliving the harms through a drawn-out legal action and potential press coverage thereof. After enduring that process, even if victorious, they may still be unsuccessful in repairing the reputational damage initially caused by the deepfake. That damage may have been done when the deepfake initially spread and people were first exposed to the video content, months (or even years) before litigation concluded.

Additionally, the same jurisdictional concerns exist as with criminal cases when the responsible party is outside the reach of United States law. Even if a plaintiff can identify the perpetrator(s), it may be impossible to hold them accountable if they are outside the jurisdiction of the United States. Finally, although a legal victory against the culpable party may provide some relief, pursuing the litigation may be a Pyrrhic victory if the defendant does not have significant financial resources. A responsible party with deep pockets, however, makes a more attractive defendant.

4. The Role of Section 230

Because deepfakes are likely to spread on social media, some might argue those platforms should bear responsibility for damage caused by such technology. However, as the law currently stands, plaintiffs injured by deepfakes will not have success going after the platform where the video content was shared. Such social platforms are shielded from immunity by Section 230 of the Communications Decency Act, which provides that users and providers of interactive computer

[133] Holly Kathleen Hall, *Deepfake Videos: When Seeing Isn't Believing*, 27 CATH. U.J.L. & TECH. 51, 69–70 (2018).

services cannot be legally treated as the publisher or speaker of third party content.[134] Unless a social media platform is itself responsible for the creation of a deepfake, this law insulates it from liability when a user posts such content on its platform. However, a plaintiff is still free to pursue the user that posted the problematic content.

Since its inception, courts have adopted a broad view of Section 230 immunity that covers a broad range of defendants and most forms of liability alleged against them.[135] In fact, Section 230 has evolved into what many commentators consider "one of the most valuable tools for protecting freedom of expression and innovation on the Internet."[136] It has immunized Facebook, Google, Yahoo!, and many others from liability stemming from third-party content whether or not the platform knew about or tried to block, remove, or police the content.[137] Courts have applied this immunity to claims for defamation, negligence, intentional infliction of emotional distress, privacy, terrorism support, and more.[138]

---

[134] 47 U.S.C. § 230(c)(1), (f)(2) (2018).
[135] *See* Mark A. Lemley, *Rationalizing Internet Safe Harbors*, 6 J. TELECOMM. & HIGH TECH. L. 101, 103 (2007) ("[Section 230] has been interpreted quite broadly to apply to any form of Internet intermediary, including employers or other companies who are not in the business of providing Internet access and even to individuals who post the content of another. And it has been uniformly held to create absolute immunity from liability for anyone who is not the author of the disputed content, even after they are made aware of the illegality of the posted material and even if they fail or refuse to remove it.").
[136] *Section 230 of the Communications Decency Act*, ELEC. FRONTIER FOUND., https://www.eff.org/issues/cda230 (last visited Jan. 28, 2018).
[137] *See generally* Fields v. Twitter, Inc., 217 F. Supp. 3d 1116, 1118 (N.D. Cal. 2016), *aff'd*, 881 F.3d 739 (9th Cir. 2018); Klayman v. Zuckerberg, 753 F.3d 1354 (D.C. Cir. 2014); Goddard v. Google, Inc., 640 F. Supp. 2d 1193 (N.D. Cal. 2009); Gentry v. eBay, Inc., 121 Cal. Rptr. 2d 703 (Ct. App. 2002); Barnes v. Yahoo!, Inc. (Barnes II), 570 F.3d 1096 (9th Cir. 2009); Carafano v. Metrosplash.com, Inc., 339 F.3d 1119 (9th Cir. 2003); Dart v. Craigslist, Inc., 665 F. Supp. 2d 961 (N.D. Ill. 2009).
[138] *See* Batzel v. Smith, 333 F.3d 1018, 1020, 1026-27 (9th Cir. 2003) (defamation); Ben Ezra, Weinstein, & Co. v. Am. Online, Inc., 206 F.3d

The rationale for such breadth was simple; it was intended to "encourage interactive computer services and users of such services to self-police the Internet for obscenity and other offensive material . . . ."[139] Before Section 230 was enacted, if a website attempted to moderate third party posts to spot and remove harmful material, it was treated as a publisher for liability purposes and could be held liable if it was unsuccessful in removing *all* such material. But a website that ignored the harm and therefore had no involvement, either in the form of publication or removal, with the content was off the hook.[140]

At the same time, Section 230 allowed companies to create platforms that relied almost entirely on user-generated content without fear of liability for the content users posted. Without the protection of this law, "the *potential* liability that would arise from allowing users to freely exchange information with one another, at this [large] scale, would have been astronomical" and could very well have prevented investors from supporting social platforms.[141] Indeed, Section 230 has been heralded as "one of the most valuable tools for protecting freedom of expression and innovation on the Internet."[142]

980, 983-84 (10th Cir. 2000) (defamation & negligence claims); Zeran v. Am. Online, Inc., 129 F.3d 327, 330, 332 (4th Cir. 1997) (negligence claims); Beyond Sys. v. Keynetics, Inc., 422 F. Supp. 2d 523, 525, 536 (D. Md. 2006) (claim under Maryland Commercial Electronic Mail Act); Doe v. Bates, No. 5:05-CV-91-DF-CMC, 2006 U.S. Dist. LEXIS 93348, at *2-*3 (E.D. Tex. Dec. 27, 2006)(involving claims of negligence, negligence per se, intentional infliction of emotional distress, invasion of privacy, civil conspiracy and distribution of child pornography); Barnes v. Yahoo!, Inc., No. 05-296-AA, 2005 U.S. Dist. LEXIS 28061, at *1 (D. Or. Nov. 8, 2005) (negligence claim resulting in personal injury).

[139] Batzel v. Smith, 333 F.3d 1018, 1028 (9th Cir. 2003); *see also* 47 U.S.C. § 230(b) (2018).

[140] Batzel v. Smith, 333 F.3d at 1029.

[141] David Post, Opinion, *A Bit of Internet History, or How Two Members of Congress Helped Create a Trillion or So Dollars of Value*, WASH. POST (Aug. 27, 2015), https://www.washingtonpost.com/news/volokh-conspiracy/wp/2015/08/27/a-bit-of-internet-history-or-how-two-members-of-congress-helped-create-a-trillion-or-so-dollars-of-value.

[142] *Section 230 of the Communications Decency Act*, *supra* note 137.

Legal scholar David Post writes that "[n]o other sentence in the U.S. Code ... has been responsible for the creation of more value"[143] than Section 230, and numerous others agree.[144]

Not everyone is as enthusiastic about Section 230, however, and whether it offers too much protection for social platforms is hotly debated by legal scholars. Many argue that it should be either revised or rescinded entirely.[145] The White House technology adviser has suggested that Congress should consider revising the law, and several lawmakers (including Senators Ted Cruz and Josh Hawley) have discussed to significantly amending or repealing Section 230.[146] Even

---

[143] Post, *supra* note 142.
[144] *See* Benjamin Edelman & Abbey Stemler, *From the Digital to the Physical: Federal Limitations on Regulating Online Marketplaces*, 56 HARV. J. ON LEGIS. 141, 160 (2019) (collecting comments).
[145] *See*, *e.g.*, Mary Graw Leary, *The Indecency and Injustice of Section 230 of the Communications Decency Act*, 41 HARV. J.L. & PUB. POL'Y 553, 557 (2018) (arguing that "although § 230 was never intended to create a regime of absolute immunity for defendant websites, a perverse interpretation of the non-sex-trafficking jurisprudence for § 230 has created a regime of de facto absolute immunity from civil liability or enforcement of state sex-trafficking laws"); Danielle Keats Citron & Benjamin Wittes, *The Problem Isn't Just Backpage: Revising Section 230 Immunity*, 2 GEO. L. TECH. REV. 453, 454 (2018) (arguing "that Section 230 immunity is too sweeping"); Andrew P. Bolson, *Flawed but Fixable: Section 230 of the Communications Decency Act at 20*, 42 RUTGERS COMPUT. & TECH. L.J. 1, 2 (2016) (offering "several proposals on how to fix some of the flaws in the existing statutory framework"); Joshua A. Geltzer, *The President and Congress Are Thinking of Changing This Important Internet Law*, SLATE (Feb. 25, 2019), https://slate.com/technology/2019/02/cda-section-230-trump-congress.html; Jeff Kosseff, *Section 230 created the internet as we know it. Don't mess with it*, L.A. TIMES (Mar. 29, 2019, 3:05 AM), https://www.latimes.com/opinion/op-ed/la-oe-kosseff-section-230-internet-20190329-story.html.
[146] Joshua A. Geltzer, *The President and Congress Are Thinking of Changing This Important Internet Law*, SLATE (Feb. 25, 2019), https://slate.com/technology/2019/02/cda-section-230-trump-congress.html; Emily Birnbaum, *Pelosi put tech on notice with warning of 'new era' in regulation*, THE HILL (Apr. 12, 2019, 1:48 PM), https://thehill.com/policy/technology/438652-pelosi-warns-its-a-new-era-for-regulating-big-tech.

Speaker Nancy Pelosi has suggested that changes could be afoot for Section 230.[147]

Despite its impact on the development of the Web 2.0, it is conceivable that Section 230 could be revised to create legal exposure for platforms on which deepfakes are spread. Consider that until last year Congress had not diminished the scope of Section 230 immunity since its inception in 1996. But in 2018 it passed the Allow States and Victims to Fight Online Sex Trafficking Act of 2017 ("FOSTA"),[148] which shrunk Section 230's civil immunity. The bill, designed to attack the online promotion of sex trafficking, carved out an exception to Section 230 for civil claims relating to sex trafficking and thus allowed those harmed to hold platforms civilly accountable.[149] It is possible that the threat of deepfakes would cause Congress to chip further at Section 230's immunity. Although this may ameliorate concerns about jurisdiction and satisfaction of judgment for plaintiffs, it could also fundamentally change the landscape of online video sharing.

5. Recent and Proposed Legislation

Seizing the idea that deepfakes have the capacity to cause serious harm, several states have considered legislation, and two have recently enacted statutes to address the concern. Texas was the first state to criminalize the creation of certain deepfakes. Its law, enacted in September 2019, makes it a criminal offense to create "a deceptive video with intent to influence the outcome of an election."[150] Although it was

---

[147] Taylor Hatmaker, *Nancy Pelosi warns tech companies that Section 230 is 'in jeopardy'*, TECHCRUNCH (Apr. 12, 2019), https://techcrunch.com/2019/04/12/nancy-pelosi-section-230/.
[148] Eric Goldman, *The Complicated Story of FOSTA and Section 230*, 17 FIRST AMEND. L. REV. 279, 280 (2018).
[149] Julio Sharp-Wasserman & Evan Mascagni, *A Federal Anti-SLAPP Law Would Make Section 230(c)(1) of the Communications Decency Act More Effective*, 17 FIRST AMEND. L. REV. 367, 400 (2019).
[150] S.B. 751, 2019 Leg., 86th Sess. (Tex. 2019), *available at* https://capitol.texas.gov/BillLookup/Text.aspx?LegSess=86R&Bill=SB751 (last visited July 10, 2019).

intentionally limited to deepfakes aimed at influencing elections in order to avoid First Amendment concerns,[151] the law is unlikely to survive strict scrutiny because it targets speech on the basis of its falsity—a notion the Supreme Court has rejected when it ruled that the Constitution protects false speech.[152] Of greater concern is that the law also suffers from vagueness and overbreadth, as its definition of unlawful deepfakes includes Constitutionally-protected expressions of parody and satire.

In October 2019, California enacted two bills to address the threat of deepfakes: AB 730 and AB 602. AB 730 criminalizes the creation or distribution of deepfakes in order to coerce or deceive voters immediately before elections,[153] and AB 602 provides a private right of action for victims of sexually explicit deepfakes.[154]

AB 730 excludes satire or parody videos but will still likely face First Amendment challenges for overbreadth because it attempts to expand the category of unprotected speech. Additionally, the California Broadcasters Association has expressed concerns that the bill would be impossible for broadcasters to comply with, and would have the unintended effect of chilling speech because broadcasters would be risk-averse to sharing even legitimate political advertising.[155]

---

[151] Lucas Ropek, *Handful of States Begin Legislating "Deepfake" Videos*, GOV'T TECH. (Apr. 30, 2019), https://www.govtech.com/policy/Handful-of-States-Begin-Legislating-Deepfake-Videos.html.
[152] United States v. Alvarez, 567 U.S. at 718.
[153] A.B. 730, 2018-2019 Leg. Sess. (Ca. 2019), *available at* https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730.
[154] A.B. 602, 2018-2019 Leg. Sess. (Ca. 2019), *available at* http://leginfo.legislature.ca.gov/faces/billCompareClient.xhtml?bill_id=201920200AB602.
[155] Luke Wachob, *California's "Deepfake" Bill is a Bad Omen,* INSTITUTE FOR FREE SPEECH (July 18, 2019), https://www.ifs.org/blog/californias-deepfake-bill-is-a-bad-omen/.

AB 602 is the first law of its kind, directed at giving private individuals recourse when their image is used for sexually explicit content without their consent. Although it may face some challenges due to vagueness, AB 602 has been narrowly drafted to target specific speech harms that are outside Constitutional protection, making its success more likely.

State legislators in New York considered a bill that would have criminalize deepfakes but only those created for trade purposes. The bill passed the New York State Assembly but expired at the end of the term while under consideration in the state senate.[156] The law would have "[e]stablish[ed] the right of privacy and the right of publicity for both living and deceased individuals" and would set up protections around "an individual's persona," defined as "the personal property of the individual [that] is freely transferable and descendible."[157] The bill attempted to accomplish two goals: to extend a postmortem right of publicity in New York and to curb the creation of unauthorized digital replicas of individuals. It included a specific provision targeted at nonconsensual pornographic deepfakes. This was likely due to the fact that New York's recently enacted law against revenge pornography, like many state laws, is unlikely to cover such nonconsensual pornographic deepfakes. Many state revenge pornography statutes are narrowly written to prohibit the dissemination of real images taken of an individual that were expected to remain private—factors that are not present when the images are computer generated and the subject is not aware of their

---

[156] Judy Bass, *New York Right of Publicity Bill Passage Drama Ends With No Action by State Senate*, N.Y. State Bar (Jun. 25, 2018) http://nysbar.com/blogs/EASL/2018/06/new_york_right_of_publicity_bi.html.

[157] N.Y. STATE ASSEMB., **A08155** (N.Y. 2018), https://nyassembly.gov/leg/?default_fld=&leg_video=&bn=A08155&term=2017&Summary=Y&Text=Y; Tom Nicholson, *Why Are Disney and Other Hollywood Giants Against 'Deepfake' Porn Legislation?,* ESQUIRE (Dec. 6, 2018), https://www.esquire.com/uk/latest-news/a21283100/why-are-disney-and-other-hollywood-giants-against-deepfake-porn-legislation/ (last visited July 10, 2019).

existence.[158] Media giants including Disney, NBC Universal, Warner Bros, Viacom, and others already roundly oppose the bill.[159] Their concern is that the breadth of the statute could inhibit those in the entertainment industry in use of computer generated characters.[160]

The first federal bill to criminalize the creation and distribution of certain harmful deepfakes was introduced by Senator Ben Sasse in late 2018.[161] The bill would make it a federal felony for individuals to:

> (1) create, with the intent to distribute, a deep fake with the intent that the distribution of the deep fake would facilitate criminal or tortious conduct under Federal, State, local, or Tribal law; or (2) distribute an audiovisual record

---

[158] *See, e.g.*, 720 ILL. COMP. STAT. ANN. 5/11-23.5 ("A person commits non-consensual dissemination of private sexual images when he or she (1) intentionally disseminates an image of another person (2) obtains the image under circumstances in which a reasonable person would know or understand that the image was to remain private; and (3) knows or should have known that the person in the image has not consented to the dissemination"; Vt. Stat. Ann. tit. 13, § 2606 (2019) ("A person violates this section if he or she knowingly discloses a visual image of an identifiable person who is nude or who is engaged in sexual conduct, without his or her consent.").

[159] Tom Nicholson, *Why Are Disney And Other Hollywood Giants Against 'Deepfake' Porn Legislation?*, ESQUIRE (Dec. 06, 2018), https://www.esquire.com/uk/latest-news/a21283100/why-are-disney-and-other-hollywood-giants-against-deepfake-porn-legislation/ (last visited July 10, 2019).

[160] *See* Letter from Lisa Pitney, Vice President of Government Relations of The Walt Disney Company, to Martin Holden, New York State Senator (June 8, 2018), *available at* https://www.rightofpublicityroadmap.com/sites/default/files/pdfs/disney_opposition_letters_a8155b.pdf; Memorandum from NBC Universal, in opposition to New York Assembly Bill A08155B, to the New York State Assembly (June 8, 2018), *available at* https://www.rightofpublicityroadmap.com/sites/default/files/pdfs/nbc_opposition_a8155b.pdf.

[161] Malicious Deep Fake Prohibition Act of 2018, **S. 3805**, 115th Cong. (2018).

> with—(A) actual knowledge that the audiovisual record is a deep fake; and (B) the intent that the distribution of the audiovisual record would facilitate criminal or tortious conduct under Federal, State, local, or Tribal law.[162]

The proposed law expired at the end of 2018, but Senator Sasse plans to reintroduce it.[163] The legislation suffers from several limitations. First, the law only targets deepfakes that are *distributed*. Creation and personal use of a deepfake is not criminalized under the statute even where it may otherwise violate the law. Only creation *with the intent to distribute* is actionable. A mere threat to release a deepfake meant to extort something of value but without distribution would thus not fall under the statute. That such conduct may be punishable under extortion laws reveals a second limitation of the statute. All of the conduct prohibited under the draft bill is already prohibited under existing law—it just further criminalizes those wrongs.

For example, as Professor Orin Kerr noted, the proposed law makes it a federal crime to make or distribute a deepfake *when the creator or distributor intends to engage in a prohibited act*.[164] However, "[i]t's already a crime to commit a crime under federal, state, local, or tribal law. It's also already a crime to 'facilitate' a crime . . . . [and] it's already a tort to commit a tort under federal, state, local, or tribal law."[165]

---

[162] *Id.*

[163] Introduced a day before the government shutdown, the bill flew under the radar and expired when the year ended. But Sasse's office reports that he intends to reintroduce it.; *See* AXIOS, *The Newest Front in the Deepfakes War: Does Congress Need to Step In?,* WWW.COUNTABLE.US (Jan. 31, 2019), https://www.countable.us/articles/20740-newest-front-deepfakes-war-does-congress-need-step (stating that Senator Mark Warner and House Intelligence Chairman Adam Schiff were also reported to be considering deepfakes legislation).

[164] Orin Kerr, *Should Congress Pass a Deepfakes Law?*, VOLOKH CONSPIRACY (Jan. 31, 2019), https://reason.com/2019/01/31/should-congress-pass-a-deep-fakes-law.

[165] *Id.*

Indeed, several criminal and civil laws are already in place that could address the harms caused by deepfakes. Senator Sasse's bill adds to that list only in that it toughens the potential punishment,[166] but it does not independently or meaningfully define and prohibit problematic content because it is limited to only that which already is legally prohibited.

A third limitation of Senator Sasse's proposed bill is the actual knowledge requirement. Assigning liability to *any* party with actual knowledge that it is distributing a deepfake recognizes the critical role individual parties will play in the spread or repression of damaging deepfakes. The infrastructure of the Internet coupled with the culture of online news and information sharing makes social platforms the most likely distribution channels for deepfakes. Social media companies will thus violate this criminal statute if they have actual knowledge that users are uploading deepfakes.[167] Though the bill does not define "actual knowledge," it presumably refers to direct and clear knowledge, something more than willful blindness or recklessness.[168]

While it seems clear that there is no onus on distributors to proactively ferret out deepfake content, it is unclear when social platforms will be considered to have actual knowledge of deepfakes on their services. This is made more difficult as effective and comprehensive detection algorithms do not currently exist. Platforms may wonder whether it will constitute actual knowledge if a single user flags a video as a deepfake. Will social platforms bear the burden of providing

---

[166] Violation of the proposed law results in fines or imprisonment of 2 years or less, except for deepfakes aimed at "affect[ing] the conduct of any administrative, legislative, or judicial proceeding of a Federal, State, local, or Tribal government agency, including the administration of an election or the conduct of foreign relations [or to] facilitate violence." Malicious Deep Fake Prohibition Act of 2018, **S. 3805**, 115th Cong. §2(a) (2018).

[167] This excepts deepfakes otherwise protected by the First Amendment.

[168] Black's Law Dictionary defines "actual knowledge" as "[d]irect and clear knowledge, as distinguished from constructive knowledge." *Actual Knowledge*, BLACK'S LAW DICTIONARY (11th Ed. 2019).

users a mechanism of reporting deepfakes separate from the current options for flagging content as problematic? Will platforms be required to respond to all content flagged even where corroborating evidence is not provided?[169] What would the required threshold then be for social platforms to respond to and remove alleged deepfakes?

An actual knowledge requirement effectively enables social platforms to bury their heads in the sand about the presence of deepfakes until they are presented with concrete evidence. By the time it is clear the content in question is a deepfake, the harm will likely already be done—a reputation may be damaged, emotional distress sustained, or public opinion swayed. In not demanding more of distributors by defining what constitutes actual knowledge and requiring proactivity in the reduction or elimination of deepfakes, the law lacks teeth.[170]

If a capable detection algorithm were developed under the law, social platforms would likely rely on such technology to reduce the possibility of criminal liability. The risk of criminal exposure would likely incentivize companies to over-police its own content. Considering the sheer volume of content uploaded and the high cost of distinguishing between

---

[169] Research shows that notice and takedown systems are routinely misused by users as a means of having content removed that they do not like or approve. *See, e.g.,* Lydia Pallas Loren, *Deterring Abuse of the Copyright Takedown Regime by Taking Misrepresentation Claims Seriously*, 46 WAKE FOREST L. REV. 745 (2011); Matthew Schonauer, *Let the Babies Dance: Strengthening Fair Use and Stifling Abuse in DMCA Notice and Takedown Procedures*, 7 J.L. & POL'Y FOR INFO. SOC'Y 135, 136 (2011).

[170] A broader knowledge requirement—one that includes recklessness—is not necessarily the answer. That would broaden criminal liability to include companies that disregard substantial and unjustifiable risks that deepfakes will be distributed on their platforms. This means that if social platforms had the ability to employ a detection algorithm capable of identifying a deepfake at the point of upload, there is a strong argument that it would be reckless not to do so, because not doing so would consciously disregard a substantial and unjustifiable risk that deepfakes will be distributed. Without a capable detection algorithm, the question of what constitutes recklessness becomes as complicated as it does for actual knowledge.

protected and actionable deepfakes, it is reasonable to project that social platforms would employ the filter and block all deepfake content[171] rather than sift through audiovisuals tagged as deepfakes and hand-identify those protected by the First Amendment.[172] However, deepfakes created for the purpose of political satire, vocal avatars, or entertainment, among myriad beneficial uses, would be blocked because it is not feasible for these platforms to mechanically differentiate between every positive and harmful use of the technology. As the Electronic Frontier Foundation notes, "platforms [already] can't tell the difference between hyperbole and hate speech, sarcasm and serious discussion, or pointing out violence versus inciting it."[173] Over-removing posts would not simply eliminate the possibility of individuals sharing positive deepfakes through their social platforms; there is also a risk it would chill positive uses of the technology at large in the process.

---

[171] YouTube's ContentID system for tracking copyrighted material serves as an example. YouTube scans uploaded videos against a database of files that have been submitted by content owners. Based on the preference of the content owner, YouTube automatically blocks the upload or permits the upload and monetizes it for the benefit of the content owner. Users blocked by this automated system can challenge the decision, but the default is to automate the blocking of the upload. *See How Content ID Works*, YOUTUBE HELP, https://support.google.com/youtube/answer/2797370 (last visited May 24, 2019).

[172] It is important to note that if distributors chose to employ content moderators for this function, unless those moderators are lawyers, identifying whether a particular video is protected by the First Amendment is not a simple task and is not one that moderators particularly enjoy. *See* Lauren Weber & Deepa Seetharaman, *The Worst Job in Technology: Staring at Human Depravity to Keep It Off Facebook*, WALL ST. J. (Dec. 27, 2017, 10:42 PM), https://www.wsj.com/articles/the-worst-job-in-technology-staring-at-human-depravity-to-keep-it-off-facebook-1514398398 [https://perma.cc/JG62-63XN].

[173] Jason Kelley & Aaron Mackey, *Don't Repeat FOSTA's Mistakes*, ELEC. FRONTIER FOUND. (Mar. 29, 2019), https://www.eff.org/deeplinks/2019/03/dont-repeat-fostas-mistakes.

In June 2019, Rep. Yvette Clark introduced the DEEPFAKES Accountability Act.[174] This bill would require mandatory watermarks and clear labeling on all deepfakes, a step that is likely to be ignored by those whose entire purpose is to weaponize a deepfake. The bill broadly defines deepfakes as any media that falsely "appears to authentically depict any speech or conduct of a person" and is produced substantially by "technical means." This expansive definition could sweep up certain protected speech particularly because the bill stumbles through its exceptions (such as entertainment and parody), failing to clarify terms and likely subjecting it to First Amendment challenges. In an Orwellian twist, the bill even exempts officers and employees of the United States who create deepfakes in furtherance of public safety or national security.[175] Like the proposed legislation by Senator Sasse and New York legislators, this bill has not progressed since its introduction in the House.

V. RECOMMENDATIONS

Deepfakes are admittedly frightening, but the government's hurried approach to regulation is as well. Legislation requires careful deliberation, especially when it is targeted at an emerging technology. This is particularly true where, as here, there are positive uses for the technology that come with strong First Amendment protections. For a legislative solution to be effective, it would need to balance these factors and account for the fact that the technology and the way it is used will continue to evolve. Shortcutting this process risks enacting laws that not only fail in their policy goals but also threaten First Amendment interests. Legal concerns around deepfakes highlight a common tension created by new technology: while there is a desire for law to keep pace with innovation to protect citizens from harmful and

---

[174] Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019, **H.R. 3230**, 116th Cong. (2019).
[175] *Id.*

unintended consequences, doing so creates a risk of stifling expression and innovation. Striking the right balance is thus critical to avoid overregulation.

Where, as here, there exists a basic framework to address many concerns associated with a new technology, there is an opportunity to observe how the law and that technology will evolve and interact as challenges emerge. This does leave room for positive innovation. It is possible that the patchwork of available laws will not adequately deter, punish, or provide sufficient restitution for the harms caused by deepfakes. This is likely to be the case with revenge pornography statutes. Although most states have enacted legislation to ban and legislate against revenge pornography, the majority of those statutes would not cover nonconsensual pornography created by deepfake technology because these statutes generally criminalize the public sharing of *actual* photos or videos of the victim. This does create an opportunity for legislators to amend statutes to ensure that they would extend to artificial representations of the victims, such as those produced by deepfake technology. California's new law, AB 602, aimed at offering victims of nonconsensual deepfakes a private right of action, is an example of bridging this gap to ensure victims have recourse.

And while it may be sensible to enact new legislation to respond to gaps such as that in revenge pornography statutes, legislators must very carefully consider proposed laws to make sure that *all* interests, including those related to the First Amendment, are considered. When legislation is quickly enacted in response to emerging technology, it can result in overregulation and other unintended negative consequences.

An example from 2015 is illustrative. That year drone legislation emerged across the country. When a trio of bills introduced in California "that would have prohibited civilians from flying aerial drones over wildfires, schools, prisons and

jails" were vetoed by then-California Governor Jerry Brown.[176] Despite alarm over privacy concerns and close calls with firefighting aircraft, Governor Brown rejected the legislation not because drone technology was without risk. Indeed, in his veto he wrote that it "certainly raises novel issues that merit careful examination."[177] Brown's veto stemmed instead from his concern that the legislation could "expose the occasional hobbyist and the FAA-approved commercial user alike to burdensome litigation and new causes of action."[178] The new laws also did not add anything to existing legal framework. They simply "multipli[ed] and particulariz[ed] criminal behavior [to] criminalize conduct that is already proscribed," which created "increasing complexity without commensurate benefit."[179] The legislation, as often happens in response to emerging technology, was rushed. "Before we go down that path," Governor Brown said, "let's look at this more carefully."[180]

A. A More Careful Look

Deepfakes pose unique challenges because they spread quickly, can be difficult to detect, and erode our conception of reality that seeing is believing. All three of these challenges are being addressed with a technological means to authenticate video content. Some sort of detection algorithm could detect deepfakes at the outset, prevent their dissemination, and make it more difficult to deny the truth of authentic videos. Drawbacks of this solution, however, are that its utilization would be voluntary, and broad application could potentially

---

[176] Patrick McGreevy, *With Strong Message Against Creating New Crimes, Gov. Brown Vetoes Drone Bills*, L.A. TIMES (Oct. 3, 2015), https://www.latimes.com/politics/la-me-pc-gov-brown-vetoes-bills-restricting-hobbyist-drones-at-fires-schools-prisons-20151003-story.html.
[177] David Siders, *Jerry Brown Vetoes Drone Regulation*, SACRAMENTO BEE (Sept. 9, 2015), https://www.sacbee.com/news/politics-government/capitol-alert/article34632729.html.
[178] *Id.*
[179] McGreevy, *supra* note 176.
[180] *Id.*

result in the over-removal of deepfakes, whether unlawful or not. As researchers work to develop detection algorithms, we must confront that, though important, they are not a panacea for problems created by deepfakes.

Likewise, legal solutions provide an important means of deterrence and restitution but alone are not a comprehensive response to the problem of deepfakes. They do little, for example, to reduce perhaps the most devastating harm caused by the technology—the potential for viewers to deny what is real by dismissing it as a deepfake. Legal solutions may have no application where the liable parties are anonymous or outside the reach of jurisdiction. Even if additional legislation becomes necessary as harmful deepfakes emerge and evolve, the more prudent course of action is to see where gaps in the law exist rather than doubling down on the laws currently in place.

Given the complexity of workable technological and legal solutions, an opportunity emerges for social platforms to lead the charge in fighting deepfakes. Driven not by legal mandates but instead by corporate citizenship and social responsibility, social platforms may be in the best position to strike the balance between supporting the growth of deepfakes for positive applications while preventing the dissemination of problematic uses.

This would be a marked change from the current climate in which social media's response to the use of its platforms for political division, fake news, terrorist recruitment, and hate speech has come too late or the pattern has been altogether ignored. Public outrage at the inaction of social platforms and Congressional inquiries regarding hate speech and political censorship on such platforms have cast them as companies that prioritize profit and avoid accountability. Thus, it is sensible for social platforms take the lead in stepping up to the challenges created by deepfakes.

Indeed, ignoring the issue of fake news costs social platforms. When users began to realize they were being

manipulated to drive profits, it resulted in a loss of trust of social platforms and Congressional inquiries into the matter. Social platforms were forced to respond. They did so by "ramp[ing] up technical efforts [and] building algorithms to 'contextualize' news with other sources on the issue. They changed their rules around fake accounts and disinformation. They hired more staff to deal with the issue."[181] Social platforms could stand to gain from meaningful efforts geared towards addressing the threats caused by deepfakes.

Although deepfakes spread quickly by a variety of users once disseminated, it all begins with a single upload. In the absence of a workable detection algorithm, how can social platforms succeed at this effort? As a first step, those platforms that have been and remain most likely to be the distribution channels for deepfakes—Facebook, Google, and Reddit, for example—should take an active role in the research to detect deepfakes. They have the financial resources and technological expertise to contribute and much to lose if user trust in their platforms continues to erode.

In the meantime, they can employ the tools they have to address this threat. Social platforms should have clear flagging procedures for fake news or fake video content, and, indeed, many already do. When content is reported fake, social platforms can temporarily remove the video while it goes to a human moderator to evaluate both whether the video is authentic and the risk of harm caused by the release of the video if it is not authentic. Crude detection algorithms—those that count blinks per minute, examine file sizes, or compare video uploads against databases of existing video content—should be utilized alongside human content moderators to determine whether the content is real or fake. By identifying the accounts where deepfakes are likely to originate typically

---

[181] Justin Sherman, *Fighting Deepfakes Will Require More Than Technology*, NEXTGOV (Dec. 14, 2018), https://www.nextgov.com/ideas/2018/12/fighting-deepfakes-will-require-more-technology/153530/.

newly created accounts or those accounts with few followers, social platforms can add additional verification filters as needed to ensure that uploaded files are authentic.

Identifying potentially deepfake content is just the first of the necessary steps. To effectively address this threat, a media literacy component is necessary. To this end, social platforms can educate users about where the information on their feed comes from. (i.e. "This comes from a trusted source.", "This does not come from a trusted source.", "ALERT: This video has been flagged as fake.") These tips should be conspicuously placed, and users should not have to search to find this information. This highlights a real issue with the efforts social platforms have undertaken thus far to "fight" the spread of fake information. Flags and alerts meant to inform users about the veracity of certain content is typically buried, only visible if users click to share the content or independently investigate.

There are several advantages to social platforms making meaningful attempts to address the threat of deepfakes instead of waiting for legal mandates from Congress. These moves would position platforms as partners in reducing unlawful and harmful false speech online rather than a business sector in need of government regulation. It would also allow social platforms the ability to control the process instead of reacting to legislative initiatives that can take months or years to pass. Alternatively, agendas driven by corporate social responsibility can be passed in a much shorter time frame and implemented quickly.[182] In addition, when the initiative is

[182] Cheryl L. Wade, *Effective Compliance with Antidiscrimination Law: Corporate Personhood, Purpose and Social Responsibility*, 74 WASH. & LEE L. REV. 1187, 1196 (2017) (noting that unlike corporate compliance with legal systems, corporate social responsibility "is almost entirely discretionary and includes things like charitable donations. . . . the something extra that companies do to be good citizens—to be responsible."); Abigail McWilliams, Donald S. Siegel & Patrick M. Wright, *Corporate Social Responsibility: Strategic Implications*, 43 J. MGMT. STUDIES 1, 18. (2006), https://doi.org/10.1111/j.1467-6486.2006.00580.x

designed by the company facilitating the transition, there are fewer operational hurdles in its execution. Importantly, this effort would likely create new business value as public perception shifts to view social platforms as leading the charge.[183] This change would likely be embraced by users and legislators alike.

VI. CONCLUSION

There are no simple solutions to the threats posed by deepfakes. It appears unlikely that a viable detection tool will be available in the near term, particularly one that can separate protected uses (satire, parody, commentary) from abuses of the technology. In the absence of a quick fix, increased education and promotion of media literacy will be important component of addressing these threats. In the meantime, new legislation may be necessary to address specific harms that are not adequately covered by existing laws, but it will have to carefully balance First Amendment interests with these harms. Like all new technologies, the pressure is on to find a solution that accounts for the fact that the technology—and likely the way it is used—will continue to evolve. In the meantime, we must accept that seeing isn't always believing.

(noting that corporate social responsibility is not always seen as being voluntary or as a moral responsibility, but is also a strategy to enhance corporate performance).

[183] Jane Nelson, *Corporate Citizenship in a Global Context,* 3 (John F. Kennedy School of Gov't, Corporate Social Responsibility Initiative, Harvard Univ., Working Paper No. 13, 2005).