Don't Let Them Fake You Out: How Artificially Mastered Videos Are Becoming the Newest Threat in the Disinformation War and What Social Media Platforms Should Do About It

Shannon Sylvester*

TABLE OF CONTENTS

* Shannon Sylvester received her J.D. from The George Washington University Law School in 2020. She would like to thank Professor Dawn Nunziato for the idea and for her support and encouragement while drafting this Article.

I. INTRODUCTION

It looked like former President Barack Obama. It sounded like former President Barack Obama. And without a second glance it fooled the best of us into thinking it *was* former President Barack Obama.[1] The "it" was a deepfake, an artificially generated video that used images and audio cloning technology to imitate the former President, making it appear as though he was saying things that he, in fact, never said.[2]

"This is a dangerous time," the fake Obama warned as he claimed, "[w]e need to be more vigilant with what we trust from the Internet."[3] While the video convincingly portrayed the former president addressing the nation, it was only because of a lack of eloquence that the video's creator Jordan Peele gave to Obama that people questioned the truth of the video.[4]

But what if Peele refused to admit that he created the video? The video, hosted on YouTube, has over 8.3 million views.[5] BuzzFeed's title of the video, "You Won't Believe What Obama Says In This Video!" followed by an emoji wink, is an obvious attempt to get people to click on the video.[6] Many people, intrigued by the title of the video, might be tempted to click on the link, and find themselves believing it was indeed former President Barack Obama saying obscenities, rather than a fake.

Believing that the former President used profanity while addressing the nation, could at a basic level, harm the President's reputation, but at a higher level, stands to do much more damage. Beyond the President's reputation, the nation's reputation could be harmed abroad. Critics of Obama could be further inflamed by the former President's offensive remarks in the video. Peele's deepfake Obama video sought to warn us of the real possibilities of disruption that could be caused by this manipulating technology. It further attempted to show that deepfakes can impair our understanding of the truth through deception, and in the hands of bad actors, can contribute to our already toxic and uncivil political discourse. The technology used to create deepfakes is advancing, and in the future, realistic deepfakes that might not be so easily debunked threaten to disrupt our already fragile democratic infrastructure.

This Article will explore the manipulative effects of deepfakes and how their truths can spread if left unchecked, significantly disrupting democracy.

---

1. *See* David Mack, *This PSA About Fake News From Barack Obama Is Not What It Appears*, BUZZFEED NEWS (April 17, 2018), https://www.buzzfeednews.com/article/davidmack/obama-fake-news-jordan-peele-psa-video-buzzfeed [https://perma.cc/6DRW-56BE].

2. *See id.*

3. *Id.*

4. *See id.* In the video, Peele, as Obama, calls President Trump a "dipshit" and argues the world is "fucked."

5. BuzzFeed Video, "You Won't Believe What Obama Says In This Video!", YOUTUBE (Apr. 17, 2018), https://www.youtube.com/watch?v=cQ54GDm1eL0 [https://perma.cc/8HNZ-KZ9E].

6. *See generally* Jessica Silbey & Woodrow Hartzog, *The Upside of Deep Fakes*, 78 MD. L. REV. 960, 964 (2019) (claiming that "eyeballs demand catchy headlines and lots of photographs").

Part I of this Article will introduce the origins of deepfakes and explain the technology and techniques that make up a deepfake. This section will also describe how deepfakes emerged on the consumer scene, and how this can have some beneficial uses—but like any technology, many negative implications as well. Part II will focus on deepfakes as catalysts to the disinformation war. As trust in the media has waned over the years, especially in the era of "fake news," public faith in the media to deliver accurate, credible news has become increasingly important. First Amendment constraints add difficultly to legislators seeking to regulate deepfakes, especially on social media sites where companies currently enjoy immunity through Section 230 of the Communications Decency Act (CDA). Social media sites are thus a prime market for deepfakes to thrive. Realizing this, some social media companies have adopted policies banning deepfakes, but they do not go far enough.

Part III of this paper will prescribe guidance on further steps social media companies should take to combat deepfakes. Social media companies should look towards helping policymakers and legislators define more clearly what deepfakes are and develop standards aimed at addressing the manipulation issues caused by deepfakes. In addition, companies should look towards adopting technological solutions. However, because social media sites are set up and run differently, there is no "one size fits all" approach when it comes to regulation and enforcement. Therefore, this paper attempts to show the problems deepfakes can cause and offer best practices that social media companies can adopt to help prevent the onslaught of damage deepfakes threaten to do if left unguarded.

II. DEEPFAKES DEFINED

Manipulated media encompasses a wide range of material, with deepfakes falling under that umbrella.[7] Deepfakes are a type of manipulated media created entirely through artificial intelligence (AI) processes.[8] "Deep" describes the "deep-learning" aspect of deepfakes, whereas "fake" refers to the fact that the video created often depicts people saying or doing things they never said or did.[9] Deepfakes should be distinguished from shallowfakes, which are also manipulated media, but manipulated through *human* intervention rather than artificial intelligence.[10]

---

7. *See* Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CAL. L. REV. 1753, 1759 (2019).

8. *See* Alex Engler, *Fighting Deepfakes when Detection Fails*, BROOKINGS (Nov. 14, 2019), https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/ [https://perma.cc/YZF6-BFFA].

9. *See* Mary Ann Franks & Ari Ezra Waldman, *Sex, Lies, and Videotape: Deep Fakes and Free Speech Delusions*, 78 MD. L. REV. 892, 893 (2019).

10. *See* Bobby Johnson, *Deepfakes Are Solvable—But Don't Forget That "Shallowfakes" Are Already Pervasive*, MIT TECH. REV. (Mar. 25, 2019); *see, e.g.*, Sarah Mervosh, *Distorted Videos of Nancy Pelosi Spread on Facebook and Twitter, Helped by Trump*, N.Y. TIMES (May 24, 2019), https://www.nytimes.com/2019/05/24/us/politics/pelosi-doctored-video.html [https://perma.cc/86JL-VEUR] (showing a shallowfake video that went viral, featuring House Speaker Nancy Pelosi appearing to slur her speech).

The technology that creates deepfakes is relatively simple to access and use.[11] But when that technology becomes more readily available on the consumer market, that raises concerns about the widespread use of deepfakes. As deepfakes enter the mainstream and grow in popularity, it is likely the technology used to create them will also become more advanced. If all it takes now is downloading an app to your phone to create a deepfake, imagine what a malicious actor could do with more sophisticated technology.[12] This section will explain the technology surrounding deepfakes and how its uses extend beyond its initial inception in Reddit threads.

A. Neural Networks and the GAN Approach

Deepfakes use deep learning, or neural network processes, known as "Generative Adversarial Networks" or GANs to function.[13] Deep learning dates back to the 1950s, when Frank Rosenblatt attempted to build a machine with a brain.[14] The idea of giving robots minds is why deep learning processes are often referred to as "neural networks."[15]

The GAN neural network process involves two networks that work against each other to produce the outcome.[16] The first network, the generator, uses a sample dataset of images to create a new image based on the sample set.[17] The second network, the discriminator, receives the new "fake" image from the generator and determines how successful the generator was at creating a plausible image.[18] If the discriminator determines the new image is inadequate and does not match up against the subject (e.g., if the mouth does not line up when the subject speaks), the discriminator sends the image back to the generator so the generator can churn out a new and improved image.[19]

The GAN method works with both images and audio clips.[20] Jose Sotelo of AI company Lyrebird, described his company's audio AI as pattern-matching.[21] The program runs by finding the uniqueness in a voice and then

11. *See* Chesney & Citron, *supra* note 7, at 1763.
12. *See, e.g.*, REFLECT, https://reflect.tech/faceswap/hot (last visited Mar. 29, 2021) [https://perma.cc/DS95-CCGV].
13. *See* Chesney & Citron, *supra* note 7, at 1761.
14. Gary Marcus, *Is "Deep Learning" a Revolution in Artificial Intelligence?*, NEW YORKER (Nov. 25, 2012), https://www.newyorker.com/news/news-desk/is-deep-learning-a-revolution-in-artificial-intelligence [https://perma.cc/GZ9Y-5GGT].
15. *Id*.
16. *Id.*
17. *See* Ian Sample, *What Are Deepfakes – And How Can You Spot Them?*, THE GUARDIAN (Jan. 13, 2020), https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them [https://perma.cc/WJL8-MN5X].
18. *Id.*
19. *See* Chesney & Citron, *supra* note 7, at 1760–61.
20. *Id.*
21. *Sleepwalkers, Truth to Power*, IHEARTRADIO (May 30, 2019), https://www.iheart.com/podcast/1119-sleepwalkers-30880104/episode/truth-to-power-45383294/ [https://perma.cc/GHU3-3436].

attempting to recreate that uniqueness.[22] While a fake audio message could be detrimental, a fake video is often times more damaging because it betrays both hearing and sight.[23]

Another way to think of the GAN approach is as a game of trickery.[24] The first machine tries to trick its adversary, the second machine, into believing the image or audio clip is legitimate.[25] If the second machine can easily spot the fake, it sends it back to the first machine to try again.[26] The first machine tries repeatedly until it can successfully trick the second machine into believing the image or audio clip is real.[27]

GANs have made their way into the consumer sphere.[28] Using the GAN approach, many companies work on their ability to create seamless deepfakes using just one video source or one photo source.[29] The results are impressive for the minimal effort it takes to create a convincing deepfake.[30] The ease of accessing and using deepfake technology for consumers has already resulted in a variety of entertaining purposes. Deepfakes have the potential to increase creative expression and even benefit the health industry, but they also have the potential to wreak havoc on individual liberty and democratic institutions.

B. From Hollywood to Handhelds

Deepfakes are relatively new to the consumer scene, but Hollywood's special effects teams have dabbled with the technology for years. The film *Forrest Gump* (1994) included an appearance by President John F. Kennedy, digitally recreated from archival video.[31] When Paul Walker died halfway through filming *Furious 7*, his brothers served as face templates to form a digital recreation of him used in the rest of the movie.[32] Even more recently, facial mapping and AI programming made actors look years younger in the

22. *See id.*; *see also* Andrew Mason, *How Imputations Work: The Research Behind Overdub*, DESCRIPT (Sept. 17, 2019), https://www.descript.com/post/how-imputations-work-the-research-behind-overdub [https://perma.cc/QF2Q-XPX7] (providing an overview of Descript company Lyrebird's audio cloning processes).

23. *Sleepwalkers*, *supra* note 21 (explaining how the host of the podcast's voice was used to create an artificial "robo" voice that was then used to prank call the host's aunt to ask for money).

24. *Id.* at 14:03.

25. *Id.*

26. *Id.* at 14:10–14:13.

27. *Id.*

28. *See generally* REFLECT, *supra* note 12.

29. *See* Colum Murphy & Zheping Huang, *Social Media Users Entranced, Concerned by Chinese Face-Swapping Deepfake App*, TIME (Sept. 4, 2019 at 10:57 AM), https://time.com/5668482/chinese-face-swap-app-zao-deep-fakes/ [https://perma.cc/4JS8-C4WF].

30. *Id.*

31. *See Pentagon's Race Against Deepfakes*, CNN BUSINESS INTERACTIVE (2019), https://www.cnn.com/interactive/2019/01/business/pentagons-race-against-deepfakes/.

32. *See* Will Knight, *The World's Top Deepfake Artist Is Wrestling With The Monster He Created*, MIT TECH. REV. (Aug. 16, 2019), https://www.technologyreview.com/s/614083/the-worlds-top-deepfake-artist-is-wrestling-with-the-monster-he-created/ [https://perma.cc/M2KL-8WZC].

Netflix film, *The Irishman*.[33] As technology improves, some hypothesize that actors will be able to license their likeness for use in television and movies without ever needing to read lines on camera.[34] Take the same technique, put it in the hands of the consumer, and suddenly the consumer becomes the director.[35] Moviegoers can make their ideal cast ensemble for their favorite movie possible with this new technology.[36] Ever wonder what Nicholas Cage would look like in *Superman*?[37] People on the Internet did, and one of the first trends of consumer-use deepfakes included putting Cage into as many movies as possible.[38] The Internet's obsession with Cage (maybe catapulted from his appearance in *Face/Off*)[39] shows the potential for consumers to embrace their creative sides as they start reimagining film.

The fascination with swapping faces helped create a market for deepfakes that consumers can create with the push of a button.[40] For instance, a popular Chinese app, Zao, allows users to upload their own photos and then superimpose their face onto a celebrity's, making the user appear to star in famous Hollywood movies.[41] The app works in seconds, and for the short amount of time used to make the video, the quality is surprisingly good.[42] Other apps such as FaceApp gained popularity when users found enjoyment making themselves age and swap genders.[43] The gaming industry is also looking to deepfakes to help make their games more attractive, allowing consumers to "play as themselves" rather than choose a character avatar.[44]

Besides their entertainment purposes, deepfakes can enrich our educational experiences and apply to the healthcare field.[45] Using a combination of GANs with virtual reality technology, prominent historical figures can appear before our very eyes. *TIME* magazine helped create an all-immersive exhibit of a depiction of Martin Luther King Jr. giving his famed "I Have a Dream" speech.[46] Health companies and researchers have also benefitted from deepfakes by using the technology to create fake brain scans with algorithms that spot tumors.[47] Just as the GAN approach helps brings the

33. *See* Angela Watercutter, *The Irishman Gets De-Aging Right – No Tracking Dots Necessary*, WIRED (May 12, 2019), https://www.wired.com/story/the-irishman-netflix-ilm-de-aging/ [https://perma.cc/LD6H-7XSR].
34. *Sleepwalkers*, *supra* note 21, at 23:00–24:00.
35. *Id.* at 26:30–29:00.
36. *Id.*
37. *Id.*
38. *Id.*
39. *Sleepwalkers*, *supra* note 21, at 27:00.
40. Murphy & Huang, *supra* note 29.
41. *Id.*
42. *Id.*
43. *Id.*
44. Knight, *supra* note 32.
45. *See* Simon Chandler, *Why Deepfakes Are a Net Positive for Humanity*, FORBES (Mar 9, 2020), https://www.forbes.com/sites/simonchandler/2020/03/09/why-deepfakes-are-a-net-positive-for-humanity/#97adbfc2f84f [https://perma.cc/Q92L-TDAL].
46. TIME: THE MARCH (2019), https://time.com/the-march/ [https://perma.cc/AS4R-EPP8].
47. *See* Chandler, *supra* note 45.

dead back to life, it can also bring life back to a patient who has lost their voice.[48]

Although deepfakes have the potential for positive applications, the majority of deepfakes circulating the web are pornographic in nature.[49] Startup Deeptrace found that pornographic deepfakes, while accounting for about 96% of deepfakes on the Internet,[50] are also disproportionately female.[51]

In December 2017, one user on Reddit posted a thread showcasing how technology made it possible to superimpose a celebrity's face onto a porn star's face, making it appear as though the celebrity was starring in a porn video.[52] In these early stages of deepfakes, the quality was poor, and it was relatively easy to distinguish the videos as fakes. However, that did not stop the harm caused by pornographic deepfakes from spreading worldwide.

In Malaysia, for example, where gay sex is illegal, a political aide was arrested following publication of a video showing him having sex with another man.[53] While the Malaysian prime minister alleged the video was a deepfake, independent experts were unable to tell if his allegation could be proved correct. [54] If the video could have been proved to be a deepfake, the political aide may not have lost his job, even though he still suffered emotional and reputational harm. But if the video was real, it presents another challenge: those accused of committing illegal acts can falsely claim manipulation of video evidence.

As deepfakes become more sophisticated and integrated into society, they present authentication challenges in a growing landscape of disinformation.[55] Not only will individuals need to be increasingly aware of fact and source checking, but they should also be wary of the prominence for plausible deniability, with public figures able to deny the credibility of a leaked video, pointing out that it might be a deepfake.[56] To combat this grim outlook, foresight is key. Educating people about deepfakes before they become technically advanced might help quell future damage from exposure to deepfakes by boosting awareness. Social media companies need to play their part in diffusing the problem of disinformation in society by adopting policies aimed at tackling manipulative media that seeks to harm. Then, there can be hope for a world where deepfakes can exist for their beneficial purposes without compromising individual liberties and democratic institutions.

---

48. *See* Sleepwalkers*, supra* note 21, at 15:20–17:00.

49. *See* Tom Simonite, *The Web Is Drowning in Deepfakes and Almost All of Them Are Porn,* WIRED (Oct. 13, 2019), https://www.wired.co.uk/article/deepfakes-porn [https://perma.cc/F8EJ-M64E].

50. *Id.*

51. Franks & Waldman, *supra* note 9, at 893–94.

52. *See* Samantha Cole, *AI-Assisted Fake Porn Is Here and We're All F**ked,* VICE: MOTHERBOARD (Dec. 11, 2017), https://www.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn [https://perma.cc/5XAB-F3E6].

53. *See* Simonite, *supra* note 49.

54. *Id.*

55. *Id.*

56. *Id.*

III. DEEPFAKES AMPLIFY THE PROBLEM OF DISINFORMATION

> "Just remember: What you're seeing and what you're reading is not what's happening."
>
> – President Donald Trump[57]

Photos and videos add credibility to stories because we trust our senses. Deepfakes force us to betray our reliable senses of hearing and sight because by their very nature, they misrepresent something real.[58] The common saying, "seeing is believing," is less true, thanks to deepfakes.

Adding to the confusion, former President Donald Trump repeatedly criticized the media's coverage of events, questioning the credibility of the press and telling his supporters not to believe what he called, the "fake news" media.[59] Fake news is not a new issue, but one that the Trump Administration and the emergence of social media sites have exacerbated.[60] Social media sites are known catalysts for causing distrust and panic with a proliferation of false information. Deepfakes, likely to infiltrate the fake news haven of social media sites, threaten to bring a new wave of confusion around trusting our sources and senses.

This section will discuss fake news generally, and how deepfakes will likely aggravate the fake news problem. Many social media companies have written their own policies to stop the spread of deepfakes, and their awareness and policies point towards a step in the right direction.

A. Disinformation Campaigns and the Difficulty in Seeking out the Truth

Social media has created a new space for political candidates to launch their campaigns and reach their supporters.[61] There is an obsession with the idea of going viral, which essentially means mass publicity.[62] Real news and

---

57. Justin Wise, *Trump: What You're Seeing in the News 'Is Not What's Happening,'* THE HILL (July 24, 2018), https://thehill.com/homenews/administration/398606-trump-what-youre-seeing-in-the-news-is-not-whats-happening-inbox-x [https://perma.cc/9PVJ-ZHF9] (reporting on President Trump giving a speech in Kansas at the Veterans of Foreign Wars National Convention).

58. *See id.*

59. *Id.*

60. *See* McKay Coppins, *The Billion-Dollar Disinformation Campaign to Reelect the President,* THE ATLANTIC (Feb. 10, 2020, 2:30 PM), https://www.theatlantic.com/magazine/archive/2020/03/the-2020-disinformation-war/605530/ [https://perma.cc/4RK5-2MKH].

61. *See* John Wihbey, *The Challenges of Democratizing News and Information: Examining Data on Social Media, Viral Patterns and Digital Influence*, *in* Shorenstein Center on Media, Politics

and Public Policy Discussion Paper Series 2 (2014) (emphasizing that social media sites boast billions of users).

62. *See id.* at 8.

fake news alike get attention due to the trending algorithms on social media sites.[63] Businesses quickly picked up on this phenomenon and started "viral marketing."[64] The explosive effect that this phenomenon promises, reaching millions of people seemingly instantly, is an attractive prospect to any marketer. But the vast reach of social media has also led to nefarious, disinformation campaigns.

1. Weaponizing Social Media

The Philippines has the highest consumption of social media worldwide.[65] Journalist Maria Ressa said that "100% of Filipinos on the Internet are on Facebook."[66] This makes the country a perfect testing site for how influential social media campaigns can be, particularly on Facebook. The Philippines has been described as "patient zero" for using disinformation campaigns to help elect their current President, Rodrigo Duterte, before similar disinformation campaigns emerged in the U.K. with Brexit and the U.S. with former President Trump's 2016 election victory.[67] Duterte is good at playing the disinformation campaign game; when the Philippines announced new election rules in 2019[68] and Facebook started rolling out fact-checking techniques, the Duterte campaign adapted, creating ways to bypass the fact-checkers.[69] Duterte's team seemed to take a page out of a 2011 Kremlin manual that views disinformation as an "invisible radiation" appearing to take effect without individuals being realized they are being acted upon.[70]

The Duterte/Kremlin campaign tactics made their way to the U.S.[71] In the 2016 U.S. presidential election, fake news was rampant on social media

---

63. *See id*.

64. *See id.* at 25.

65. *See* GLOBAL WEB INDEX, SOCIAL 20 (2018), https://www.globalwebindex.com/hubfs/Downloads/Social-H2-2018-report.pdf [https://perma.cc/B6XD-GHMC] (finding that Filipinos spend on average 4 hours on social media a day).

66. *See* Ailsa Chang, *'A Thousand Cuts' Documentary Tracks Disinformation in Duterte's Philippines*, NPR (Feb. 3, 2020), https://www.npr.org/2020/02/03/802392333/a-thousand-cuts-documentary-tracks-disinformation-in-dutertes-philippines [https://perma.cc/7WDA-VJDT].

67. *See id*.; *see also* Craig Silverman, *The Philippines Was a Test of Facebook's New Approach to Countering Disinformation. Things Got Worse.,* BUZZFEED NEWS (Aug. 7, 2019), https://www.buzzfeednews.com/article/craigsilverman/2020-philippines-disinformation [https://perma.cc/LT5H-QPAZ] (citing an interview with Facebook's public policy director for global elections, Katie Harbarth, in which Harbath referred to the Philippines as "patient zero").

68. *See* Michael Bueza*, #PHVote: Campaign Rules for 2019 Midterm Elect*ions, RAPPLER (Feb. 28, 2019), https://www.rappler.com/nation/politics/elections/2019/224390-comelec-campaign-rules [https://perma.cc/NWC4-Y5K5].

69. *See* Silverman, *supra* note 67 (for example, Duterte's campaign avoided fact checkers by relying on microtargeting and promoting articles with minimal amounts of truth to avoid being flagged as false).

70. *See* Coppins, *supra* note 60.

71. *See id.* (noting that the Trump campaign understood the power of using "disinformation architecture" like that used in the Duterte campaign on Facebook and "methods of disinformation" referenced in a "2011 manual for Russian civil servants").

sites, with microtargeting being one of the key strategies used by candidates.[72] From #pizzagate to Pope Francis endorsing President Trump, the 2016 campaign trail was filled with falsities.[73] Misinformation drowned out fact. Even when the fake information was debunked, many social media users were already convinced. This phenomenon occurs due to the illusory truth effect.[74] The illusory truth effect describes how repeat exposure to false information increases the chances of people accepting the false information as true.[75] Ideas like counter-speech likely won't work because people who repeatedly encounter a fake story are more likely to remember it as true.[76]

In a society where we question everything, making the truth harder to discern, deepfakes will only add more uncertainty to the mix. Our continuous questioning leads us as a democratic society to value seeking out the truth, something that misleading speech carried in the medium of video manipulation makes quite difficult.[77] A Pew Research Center study conducted in November and December 2018 found that over half of the people surveyed believed that Americans' trust in the federal government and each other has been shrinking.[78] In a separate, further inquiry, around 49% of technology experts believed that technology will have a negative impact and mostly weaken core aspects of democracy, such as trust in government, in the coming decade.[79]

The decline in trust and lack of gatekeeping has made it extremely difficult to control the spread of disinformation.[80] Adding to this challenge are First Amendment concerns and platform liability issues related to Section 230 of the CDA.

2. Fake Speech is (Mostly) Free Speech

> Congress shall make no law respecting an establishment of religion or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press, or the

72. *Id.*

73. *See* Hannah Ritchie, *Read All About It: The Biggest Fake News Stories of 2016*, CNBC (Dec. 30, 2016, 2:04 AM), https://www.cnbc.com/2016/12/30/read-all-about-it-the-biggest-fake-news-stories-of-2016.html [https://perma.cc/HZ56-68QP].

74. *See* Franks & Waldman, supra note 9, at 894.

75. *See id.*

76. *Id.*

77. *Id.*

78. LEE RAINIE ET AL., PEW RES. CTR., TRUST AND DISTRUST IN AM. 3 (2019) (showing also that people believe it is important to fix this decline in trust and that the low trust makes it harder to solve problems in the U.S.).

79. Janna Anderson & Lee Rainie, *Many Tech Experts Say Digital Disruption Will Hurt Democracy*, PEW RES. CTR. (Feb. 21, 2020), https://www.pewresearch.org/internet/2020/02/21/many-tech-experts-say-digital-disruption-will-hurt-democracy/ [https://perma.cc/EB8P-JAEC].

80. Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CAL. L. REV. 1753, 1763–65 (2019).

> right of the people peaceably to assemble, and to petition the Government for a redress of grievances.[81]

Freedom of speech is a fundamental right enshrined in the U.S. Constitution. Unlike other countries, the U.S. holds free speech to such a high standard it is difficult to restrict. Fake speech is merely a consequence of allowing people to engage in civil discourse and further the belief in the "marketplace of ideas."[82] The marketplace of ideas invokes the optimism that the truth will win out eventually, but as the illusory truth effect demonstrates, the marketplace of ideas is not an immaculate concept. Deepfakes are not ideas simply countered with "better" ideas. Since deepfakes involve freedom of expression, as long as they do not end up causing physical harm to a person, laws banning or restricting deepfakes would be unlikely to pass the strict scrutiny test of the First Amendment.[83]

Deepfakes by their very nature promote fake speech, but fake speech is constitutionally protected under *New York Times v. Sullivan*.[84] In that case, the U.S. Supreme Court held that public officials cannot sue for defamation unless they prove "actual malice," meaning the plaintiff must show that the false statement was made with knowledge of its falsity and in reckless disregard to the truth.[85] The Court then rationalized in *U.S. v. Alvarez* that fake speech should be protected because by itself, fake speech can be valuable in encouraging public discourse and it does not cause any legally cognizable harm.[86]

The Supreme Court has offered little guidance when it comes to fake speech in virtual applications, like videos. But in *Ashcroft v. Free Speech Coalition*, the Court suggested in dictum that "computer morphing" (using real images of children to frame the images being used in videos) might not be protected speech because it would cause harm similar to that in an appropriation suit, using the real child's likeness without their consent.[87] Defining true harm when it comes to speech is challenging, and the Court in *Ashcroft* chose to characterize harmful speech based on its emotional and reputational impacts.[88] But the Court in *Brandenburg v. Ohio* held that speech amounts to harmful incitement only when it is likely to produce imminent lawless action, commonly referred to as the *Brandenburg* test.[89]

---

81. U.S. CONST. amend. I.
82. Franks & Weldman, *supra* note 9, at 894.
83. Chesney & Citron, *supra* note 7, at 1790.
84. *See* N.Y. Times v. Sullivan, 376 U.S. 254 (1964).
85. *Id.* at 276.
86. *See* U.S. v. Alvarez, 567 U.S. 709, 719 (2012).
87. Ashcroft v. Free Speech Coalition, 535 U.S. 234, 242 (2002) (holding that virtual child pornography is a protected form of free speech because no children are harmed and fake images are used).
88. *Id.*
89. *See* Brandenburg v. Ohio, 395 U.S. 44 (1969)

Attempts to regulate free speech, especially political speech, are often seen as an overreach of government power.[90] The fear is that speech regulation could turn partisan, with the government choosing to strike down speech with which it disagrees.[91] Some states have enacted laws banning political deepfakes, but these laws are wrought with First Amendment concerns.

a. Political Speech Deepfakes

In 2019, Texas amended its Election Code, making it a crime to create and publish a deepfake video within 30 days of an election with the intent to injure a candidate or influence the results of an election.[92] Violations of the law are punishable by up to a year in jail and a $4,000 fine.[93] The Act defines a deepfake as "a video, created with the intent to deceive, that appears to depict a real person performing an action that did not occur in reality."[94] This definition is overbroad and appears to apply to shallowfakes and deepfakes alike as it does not draw a distinction between artificially made videos. There is also no exception for satire or parody videos that have been used in campaigns before, making illegal any video that "intends to deceive."[95]

The Texas law is an example of good intentions through misguided efforts. In an attempt to ban all political deepfake videos, the Act may do more harm than good. The law threatens to define truth, inserting government as a mediator to decide what deception means. It is also so broad that numerous political ads of past and present would likely trigger criminal liability for their creators or distributors if they ran within 30 days of an election.[96] It is unlikely that this law will withstand First Amendment challenges, as it does not seem narrowly tailored enough to restrict speech.[97]

The California legislature also recently addressed deepfakes. Effective as of January 1, 2020, California Assembly Bills 602 (AB 602) and 730 (AB 730) aim to curb the distribution of deepfakes.[98]AB 602 creates a private right

90. *See* Brown v. Hartlage, 456 U.S. 45, 46 (1982) ("The State's fear that voters might make an ill-advised choice does not provide the State with a compelling justification for limiting speech.").

91. *See* Helen Norton, *Lies and the Constitution*, 2012 SUP. CT. REV. 161, 199 (2012).

92. TEX. ELEC. CODE tit. 15, § 225.004 (2019).

93. *See* Matthew F. Farraro et al., *First Federal Legislation on Deepfakes Signed into Law*, WILMERHALE (Dec. 23, 2019), https://www.wilmerhale.com/en/insights/client-alerts/20191223-first-federal-legislation-on-deepfakes-signed-into-law [https://perma.cc/87AF-XBYU].

94. TEX. ELEC. CODE tit. 15, § 225.004.

95. *Id.*

96. *See* Mark Rumold, *Not a Hoax: The Very Real Threat of Political 'Deepfakes' Laws*, ELEC. FRONTIER FOUND. (Apr. 27, 2020), https://www.eff.org/deeplinks/2020/04/not-hoax-very-real-threat-political-deepfakes-laws [https://perma.cc/YG83-VR6Q].

97. *See, e.g.*, Susan B. Anthony List v. Driehaus, 814 F.3d 466 (6th Cir. 2016) (holding election-based lies were an insufficient reason to restrict speech).

98. *See* K.C. Halm et. al, *Two New California Laws Tackle Deepfake Videos in Politics and Porn*, DAVIS WRIGHT TREMAINE LLP (Oct. 14, 2019), https://www.dwt.com/insights/2019/10/california-deepfakes-law [https://perma.cc/M4GD-SGCP].

of action for individuals depicted in sexually explicit material through digital or electronic technology.[99] Individuals can recover damages for emotional distress or statutory damages up to $150,000 if the act was committed with malice.[100]

AB 730 makes it illegal to create or distribute videos, images, or audio of politicians appearing in "fake videos" within 60 days of an election.[101] AB 730 defines "materially deceptive audio or visual media" as media involving a candidate that is intentionally manipulated and would reasonably confuse a person as to the authenticity of the media.[102] The law does not apply to satire or parody and allows fake video or audio ads as long as there is a disclosure on the video clarifying the video is manipulated.[103] AB 730 has drawn criticism for lacking First Amendment exemptions.[104] Because political speech has robust First Amendment protection, this law is likely going to be difficult to enforce.[105]

Other states are following suit. Maine, Maryland, and Washington are among those states that have proposed deepfake bills. [106] However, because of the difficulty of regulating speech, specifically political speech, it is unlikely that such bills would withstand First Amendment challenges, unless they carefully carve out First Amendment protections.[107] Political attack ads have existed for centuries, so the addition of deepfakes purporting to show candidates saying and doing things they never said or did would have to be significantly distinguished from other forms of political protected speech.

b. Defamation Actions

Deepfakes largely involve using another person's likeness without their consent, leading some to believe the remedy to combat fake speech

---

99. CAL. CIV. CODE § 1708.86 (2019).
100. *Id.*
101. CAL. ELEC. CODE § 20010 (2019).
102. *Id*.
103. *Id.*
104. *See* Evan Symon, *'Deepfake' Videos of Political Candidates in Ads Now Illegal In California,* CAL. GLOBE (Oct. 7, 2019, 8:17PM), https://californiaglobe.com/section-2/deepfake-videos-of-political-candidates-in-ads-now-illegal-in-california/ [https://perma.cc/D654-T7EW].
105. *See* Kari Paul, *California Makes 'Deepfake' Videos Illegal, But Law May Be Hard to Enforce*, GUARDIAN (Oct. 7, 2019), https://www.theguardian.com/us-news/2019/oct/07/california-makes-deepfake-videos-illegal-but-law-may-be-hard-to-enforce [https://perma.cc/6ZK5-2HEF].
106. *See* Scott Thistle, *Maine Lawmakers Take Up Bill to Ban 'Deepfake' Political Ads*, PRESSHERALD (Jan. 29, 2020), https://www.pressherald.com/2020/01/29/maine-lawmakers-take-up-bill-to-ban-deepfake-political-ads/# [https://perma.cc/3LQG-BAEN] (the proposed Maine law would prohibit the publication and distribution of a deepfake video of a candidate within 60 days of an election. The political candidate targeted in the fake ad could seek redress through a court order to block the content and the opportunity to pursue civil action against the maker of the deepfake); *see also* Matthew Feeney, *Deepfake Laws Risk Creating More Problems Than They Solve*, CATO (Mar. 1, 2021) https://www.cato.org/sites/cato.org/files/2021-03/Paper-Deepfake-Laws-Risk-Creating-More-Problems-Than-They-Solve.pdf.
107. *See* Rumold, *supra* note 96.

found in deepfakes already exists in defamation law.[108] The problem with defamation suits for combatting deepfakes is that they require a higher standard for public officials to prove the falsity of a statement. As seen in *New York Times v. Sullivan*, the burden is on the public official to prove speech is false under the actual malice standard by clear and convincing evidence.[109] The defendant, on the other hand, need not show the speech is true.[110] Proving actual malice is in theory difficult because it requires showing that the defendant had actual knowledge that the speech was false or that the person acted in reckless disregard of the truth.[111]

Getting legal remedies for a deepfake action in general might be difficult. To succeed, the plaintiff would need to know the creator of the deepfake.[112] Lawsuits are also often costly and time-consuming.[113] In addition, not all deepfakes involve a specific individual, meaning there might not be standing to sue in some cases.[114] And then there is the difficulty of suing the platform hosting the video because of CDA § 230 protections.[115]

c. CDA Section 230 Protections[116]

> No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.[117]

In the 1997 Fourth Circuit case, *Zeran v. AOL*, the court held that liability upon notice has a chilling effect on the freedom of Internet speech.[118] Since 1997, Internet speech has skyrocketed. Today, online social media platforms would be extremely burdened if they were liable to be sued for every false speech represented or posted on their site. Similarly, if such platforms became aware of fake, or what some might deem harmful posts, it would not be incumbent upon them to take down the speech, because that would contradict the marketplace of ideas theory heralded by free speech enthusiasts and Internet users alike. However, this should not give social

---

108. *See* DAVID GREENE, WE DON'T NEED NEW LAWS FOR FAKED VIDEOS, WE ALREADY HAVE THEM (2018), https://www.eff.org/deeplinks/2018/02/we-dont-need-new-laws-faked-videos-we-already-have-them [https://perma.cc/7RCA-N7NN].

109. Bose Corp. v. Consumers Union of U.S., Inc., 466 U.S. 485, 511–12 (1984); *see generally Sullivan*, 376 U.S. 254 (1964).

110. *Sullivan*, 376 U.S. at 279.

111. *See* Greene, *supra* note 108.

112. *See generally* Chesney & Citron, *supra* note 7, at 1792.

113. *Id.*

114. *Id.*

115. *Id.* at 1796.

116. This Article refers to Section 230 as part of the CDA for ease of reference, but Section 230, per the FCC, is technically part of the Communications Act. *See* Thomas M. Johnson Jr., *The FCC's Authority to Interpret Section 230 of the Communications Act*, FCC (Oct. 21, 2020), https://www.fcc.gov/news-events/blog/2020/10/21/fccs-authority-interpret-section-230-communications-act [https://perma.cc/K674-PQVM].

117. 47 U.S.C. § 230(c)(1) (2018).

118. Zeran v. AOL, 129 F.3d 327 (4th Cir. 1997).

media platforms the excuse to turn a blind eye to illegal activities occurring on their sites.[119]Although they currently have no legal obligation to monitor speech activities on their sites, society urges them to have an ethical duty to reward good behavior and encourage the free flow of ideas in a way that benefits others. Noting this, many social media companies have enacted their own Rules of Engagement or Community Standards and Polices that users must abide by if they wish to participate on those platforms. With this in mind, social media companies may not need legal repercussions to get them to act. Rather, moral and political pressures might be enough to incentivize social media companies to engage in a form of beneficial content moderation.[120]

B. What Social Media Companies are Doing About Deepfakes

In July 2019, Representative Adam Schiff (D-CA), head of the House Intelligence Committee, sent letters to social media companies asking them to describe their plans for combatting the spread of deepfakes on their sites, especially ahead of the 2020 presidential election and the growing threat of disinformation.[121] Schiff expressed (valid) concern for the proliferation of false information and misrepresentations to spread on social media sites, causing panic and distrust.[122]

Social media platforms, catalysts for wreaking havoc by spreading false information, should take steps to stop the spread of harmful, false information caused by manipulated media.[123] That includes adopting policies that: (1) define manipulated media such as deepfakes; (2) address criteria for take-down techniques; (3) comply with the First Amendment; and (4) identify the differences, if any, between political and commercial speech portrayed through manipulated media. Most of the policies currently in place fail to address at least one of these proposals.

119. *See* Chesney & Citron, *supra* note 7, at 1797.

120. *Id.* at 1795.

121. Press Release, Adam Schiff, Member, U.S. House of Representatives, Schiff Presses Facebook, Google and Twitter for Policies on Deepfakes Ahead of 2020 Election (July 15, 2019), https://schiff.house.gov/news/press-releases/schiff-presses-facebook-google-and-twitter-for-policies-on-deepfakes-ahead-of-2020-election [https://perma.cc/L7CQ-J7L9].

122. *See, e.g.*, Hugh Langley, *Rep. Adam Schiff Told Google and Twitter to Step Up Their Fight Against Coronavirus Misinformation with an Unexpected Message: Be More Like Facebook*, BUS. INSIDER (Apr. 30, 2020, 6:59PM), https://www.businessinsider.com/adam-schiff-tells-google-and-twitter-to-look-to-facebook-2020-4 [https://perma.cc/G7HU-C9BE].

123. *See, e.g.*, Jesselyn Cook, *Online Anti-Vax Communities Have Become A Pipeline for QAnon Radicalization*, HUFFINGTON POST (Nov. 28, 2020), https://www.huffpost.com/entry/qanon-anti-vax-coronavirus_n_5fbeb0c0c5b61d04bfa6921a [https://perma.cc/N22C-LL7B].

1. Facebook

Facebook claims its key to tackling harmful deepfakes is "collaboration."[124] On January 6, 2020, Monika Bickert, Facebook's Vice President for Global Policy Management, released Facebook's strategy for combatting deepfakes and other forms of manipulated media.[125] The strategy involved working with academia, government, and industry to develop solutions, as well as implementing investigations of AI-generated content.[126] Facebook, along with Amazon Web Services, Microsoft and other partners, launched the Deepfake Detection Challenge (DFDC) in September 2019.[127] The goal of the DFDC, is to bring academics and researchers together to find innovative ways to detect deepfakes.[128] Facebook also partnered with Reuters to help journalists identify deepfakes in a free online course.[129]

Facebook also has a policy in its Community Standards specifically related to manipulated media.[130] That policy states that Facebook will remove deceptive manipulated media if it has been edited or synthesized in ways such that an average person would be misled as to the authenticity of the media.[131] The policy also carves out an exception for satire or parody media.[132] Facebook's manipulated media policy has been criticized both as overbroad and too narrow.[133] As Whitney Phillips from *WIRED* put it, the policy is "best described as a slice of Swiss cheese that's mostly holes."[134]

In an attempt to avoid overregulation while still protecting free speech, Facebook allows users who have content taken down for violating Facebook's policies to challenge their takedown with an independent third-party fact checker.[135] Facebook also said it will not invariably take down manipulated media that violates its policies and will instead label the affected media.[136] Facebook argues this labelling process will help educate people as to what "fake news" is, but it is unlikely that simple labelling measures will keep

124. *See* Monika Bickert, *Enforcing Against Manipulated Media*, FACEBOOK (Jan. 6, 2020), https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/ [https://perma.cc/ZV2U-29CT].

125. *Id.*

126. *Id.*

127. *Id.*

128. *See Description of Deepfake Detection Challenge*, KAGGLE, https://www.kaggle.com/c/deepfake-detection-challenge/overview/description (last visited Feb. 15, 2021) [https://perma.cc/4C7M-XXZM?type=image].

129. *See* Reuters Staff, *Reuters Partners with Facebook Journalism Project to Help Newsrooms Around the World Spot Deepfakes and Manipulated Media*, REUTERS (Dec. 17, 2017), https://www.reuters.com/manipulatedmedia/en/ [https://perma.cc/6NU3-5WS6].

130. *Facebook Community Standards, "Manipulated Media"*, FACEBOOK (2020), https://www.facebook.com/communitystandards/manipulated_media [https://perma.cc/RWB2-QDA6].

131. *Id.*

132. *Id.*

133. *See* Whitney Phillips, *The Internet Is a Toxic Hellscape—But We Can Fix It,* WIRED, (Feb. 3, 2020), https://www.wired.com/story/the-internet-is-a-toxic-hellscape-but-we-can-fix-it/ [https://perma.cc/6DJC-39S5].

134. *Id.*

135. Bickert, *supra* note 124.

136. *Id.*

people from seeing and believing the media as true.[137] As we saw with the illusory truth effect, the opposite is in fact true.[138]

2. Twitter

About a month after Facebook announced its deepfake policy, on February 4, 2020, Twitter announced a new policy related to "synthetic and manipulated media."[139] Twitter's policy was user-focused. For example, the company posted a survey in the fall of 2019 soliciting feedback on its proposed policy from Twitter users who commented with the hashtag #TwitterPolicyFeedback.[140] After receiving around 6,500 responses globally, Twitter posted its findings and crafted its new rule.[141] The new rule states: "You may not deceptively share synthetic or manipulated media that are likely to cause harm. In addition, we may label Tweets containing synthetic and manipulated media to help people understand their authenticity and to provide additional context."[142]

Twitter's approach includes labeling deceptively altered or fabricated content, only removing the content if it impacts public safety or is likely to cause serious harm.[143]

Is the content significantly and deceptively altered or fabricated?	Is the content shared in a deceptive manner?	Is the content likely to impact public safety or cause serious harm?	
✓	✗	✗	Content **may** be labeled.
✗	✓	✗	Content **may** be labeled.
✓	✗	✓	Content is **likely** to be labeled, or **may** be removed.*
✓	✓	✗	Content is **likely** to be labeled.
✓	✓	✓	Content is **likely** to be removed.

Twitter's policy seems to apply to shallowfakes as well as deepfakes, stating that the Twitter team is likely to act on significant forms of alteration

137. *See* Donie O'Sullivan & Marshall Cohen, *Facebook Begins Labeling, but Not Fact-Checking, Posts From Trump And Biden*, CNN BUS. (July 21, 2020), https://www.cnn.com/2020/07/21/tech/facebook-label-trump-biden/index.html [https://perma.cc/29NW-EZMJ].

138. *See* Franks & Waldman, *supra* note 9, at 894–95.

139. *See* Yoel Roth & Ashita Achuthan, *Building Rules in Public: Our Approach to Synthetic & Manipulated Media*, TWITTER BLOG (Feb. 4, 2020), https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html [https://perma.cc/K6EW-XYBB].

140. *Id.*

141. *Id.*

142. *General Guidelines and Policies: Synthetic and Manipulated Media Policy*, TWITTER HELP CTR., https://help.twitter.com/en/rules-and-policies/manipulated-media (last visited May 4, 2020) [https://perma.cc/3G7C-UJZK].

143. *Id.*

such as audio or video content doctored to change its meaning.[144] This gives Twitter discretion to determine if a video is manipulated in such a way that is inauthentic to merit labels or removal from its site. Twitter maintains it will be an impartial editor, only labeling or removing videos identified by its technology or reported by a third party.[145] Some of the serious harms that could be cause for removal include threats to the privacy or ability of a person or group to freely express themselves or participate in civic events.[146]

Twitter's first case in applying its new policy encountered problems. White House social media director for former President Donald Trump, Dan Scavino, tweeted a manipulated video of (then) former Vice President Joe Biden appearing to endorse Trump for reelection in 2020, which Trump also retweeted.[147] Twitter labeled the tweet as "manipulated media," but the tag only appeared if the tweet showed up on someone's timeline and was not visible to users who tried to search for the video or physically clicked on the video.[148]

Since then, Twitter has been on a labeling frenzy,[149] going so far as to kick Trump off the site in 2021 following an attack by his supporters on the U.S. Capitol.[150] Twitter claimed it was permanently suspending Trump's account due to risk of "further incitement of violence."[151] Prior to the ban, Twitter had already started to label a slew of Trump's tweets, hiding the tweets, and limiting replies, based on Trump's false claims that he won the election and allegations of voter fraud.[152] Twitter's labeling stated: "Some or all of the content shared in this Tweet is disputed and might be misleading about an election or other civic process."[153]

3. Google/YouTube

YouTube, owned by Google, reiterated its stance on election-related content in an official YouTube blog on February 3, 2020.[154] YouTube's

144. *Id.*
145. *See id.*
146. *Id.*
147. *See* Ivan Metha, *Trump's Retweet with Doctored Biden Video Earns Twitter's First 'Manipulated Media' Label*, THE NEXT WEB (March 9, 2020), https://thenextweb.com/twitter/2020/03/09/trumps-tweet-with-doctored-biden-video-earns-twitters-first-manipulated-media-label/ [https://perma.cc/9RVZ-RWQ9].
148. *Id.*
149. *See, e.g.*, Twitter Safety, *Updates to Our Work on COVID-19 Vaccine Misinformation*, TWITTER BLOG (Mar. 1, 2021), https://blog.twitter.com/en_us/topics/company/2021/updates-to-our-work-on-covid-19-vaccine-misinformation.html [https://perma.cc/7UCK-N7X8].
150. Twitter Inc., *Permanent Suspension of @realDonaldTrump*, TWITTER BLOG (Jan. 8, 2021), https://blog.twitter.com/en_us/topics/company/2020/suspension.html [https://perma.cc/J9AX-3BQH].
151. *Id.*
152. *Id.*
153. *Id.*
154. *See* Leslie Miller, *How YouTube Supports Elections*, YOUTUBE OFF. BLOG (Feb. 3, 2020), https://youtube.googleblog.com/2020/02/how-youtube-supports-elections.html [https://perma.cc/9NTF-KR5Q].

deceptive practices policies state that: "[C]ontent that has been technically manipulated or doctored in a way that misleads users (beyond clips taken out of context) and may pose a serious risk of egregious harm" will be removed.[155] YouTube further states it will remove content that attempts to mislead people about the voting process or any other false information relating to elections.[156]

YouTube will not only remove false content if it fits the criteria, but it will also terminate channels that "[a]ttempt to impersonate another person or channel, misrepresent their country of origin, or conceal their association with a government actor."[157]

In 2018, YouTube created an Intelligence Desk to help review technically-manipulated content and take proactive approaches to mitigate the spread of the content.[158] YouTube also changed its recommendations system to prevent people from viewing misinformation on its site.[159] The Intelligence Desk and recommendation system are attempts by YouTube to be proactive and get ahead of videos before they become viral, when they can do the most damage.[160] To achieve this, YouTube relies on Google data, user reports, social media trends, and third-party consultants.[161] YouTube later added human vetting and content moderators.[162]

Google has tried to warn about the dangers surrounding deepfakes by releasing an open-source database containing 3,000 manipulated videos.[163] Google's hope was that researchers would start to develop deepfake detection tools.[164]

Also noteworthy is that YouTube found that it was within its policies to take down a shallowfake video of Nancy Pelosi appearing to slur her words during a speech.[165] Facebook, on the other hand, kept the video up.[166]

---

155. *Id.*
156. *Id.*
157. *Id.*
158. *Id.*
159. Miller, *supra* note 154.
160. *See id.*
161. Alex Kantrowitz, *YouTube Is Assembling New Teams to Spot Inappropriate Content Early*, BUZZFEED (Jan 19, 2018), https://www.buzzfeednews.com/article/alexkantrowitz/youtube-intelligence-desk-will-spot-inappropriate-content [https://perma.cc/6MAH-6BQL].
162. *Id.*
163. Karen Hao, *Google Has Released A Giant Database of Deepfakes to Help Fight Deepfakes*, MIT TECH. REV. (Sept. 25, 2019), https://www.technologyreview.com/f/614426/google-has-released-a-giant-database-of-deepfakes-to-help-fight-deepfakes/ [https://perma.cc/ME9Z-MUWH]; *see also* Nick Dufour & Andrew Gully, *Contributing to Deepfake Detection Research*, GOOGLE AI BLOG (Sept. 24, 2019), https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html[https://perma.cc/7JCW-5CTK].
164. *See* Dufour & Gully, *supra* note 163.
165. *See* Mervosh, *supra* note 10.
166. *Id.*

IV. MITIGATING THE DEEPFAKE THREAT

> "[W]e must reject a culture in which facts themselves are manipulated and even manufactured."
>
> – President Joe Biden[167]

Trusting the authority of public officials and the government generally will play a huge role in helping to combat the threat of deepfakes. Instead of fostering distrust in the media, the Biden Administration seeks to bring truth back to light. But it cannot do so alone. Social media companies have taken steps in the right direction by raising awareness of deepfakes by creating policies banning certain kinds of manipulated media from their sites.[168] However, because social media companies are largely self-regulating, their policies differ in how deepfakes are defined, and they fail to adequately protect free speech rights.[169] To provide a stronger, more united front on behalf of social media companies, proposals range from amending CDA Section 230 to investing in various technological solutions. However, perhaps the biggest challenge social media companies face in regulating deepfakes and other fake news is moderating content in line with free speech. If social media companies have too much power to regulate what is being said on their platforms, this could seriously diminish individuals' freedom of expression.

A. Amending CDA Section 230

While some have criticized amending Section 230, believing that it is vital to the Internet's existence, Danielle Citron and Benjamin Wittes are convinced that an amendment, while retaining much of platforms' liability, is viable.[170] Citron's and Wittes' proposed amendment is more of a compromise, requiring companies to use reasonable content moderation practices to earn the immunity provided by Section 230.[171] It is not impossible to amend Section 230, and the Allow States and Victims to Fight Online Sex Trafficking Act (FOSTA), which allowed greater regulation of sex trafficking content on the Internet, is proof of that.[172]

Section 230 is outdated. One of the biggest selling points of Section 230 is that it lets platforms off the hook from sifting through massive amounts of

---

167. Joseph R. Biden, Jr., President, United States of American, Inaugural Address (Jan. 20, 2021), https://www.whitehouse.gov/briefing-room/speeches-remarks/2021/01/20/inaugural-address-by-president-joseph-r-biden-jr/ [https://perma.cc/6MTY-8FUE].

168. *See supra* Section II.B.

169. *Id.*

170. *See* Chesney & Citron, *supra* note 7, at 1798–99.

171. *See* Danielle Citron & Quinta Jurecic, *Platform Justice:Content Moderation at an Inflection Point*, Hoover Inst. Essay 2 (2018), https://www.scribd.com/document/387911222/Platform-Justice-Content-Moderation-at-an-Inflection-Point#download&from_embed [https://perma.cc/8G7K-Z7KU].

172. *See* Chesney & Citron, *supra* note 7, at 1798–99.

data that would otherwise be deemed impossible to monitor.[173] However, technology, like the Internet, has evolved since then. Many sites now have compliance monitors built in, through machine learning and AI, that allows social media platforms to track and take down harmful content, such as child pornography and IP violations.[174] This approach can be applied to deepfakes as well. Social media companies already have the technology to combat the spread of harmful information on their sites, now they just need a legislative push.

With the recognized harms of deepfakes, Section 230 should and can be amended to prevent harmful disinformation from rampantly spreading on social media sites. It has been done before, and it can be done again.[175] Any proposed amendment would have to ensure social media platforms are not engaging in over-regulation and would consider the First Amendment.[176] Because false speech is not unconstitutional, an amendment to Section 230 would have to specifically account for false speech that harms. In defining speech that harms, legislators should look towards defamation actions and other appropriation torts. Congress can incentivize platforms to take down such false, harmful speech, by still granting overreaching immunity for most content published on social media sites due to the broad scope of Section 230. That is the beauty of amending, rather than dismantling and getting rid of Section 230 altogether.

B. Stronger Deepfake Legislation

Instead of placing the burden on social media platforms to monitor and remove deepfakes or face liability under a newly amended Section 230, another approach Congress could take would be to enact a federal law could successfully regulate deepfakes by clearly defining them as manipulated media. This will enable social media platforms to adapt their policies to that definition while alleviating First Amendment concerns. Most of the laws currently surrounding deepfakes in the U.S. are more research-focused[177] or related more directly to pornographic deepfakes.[178] Deepfake laws that purport to ban deepfakes for deceptive speech are largely nonexistent, likely due to concerns that such laws impermissibly block free speech.

---

173. *See supra* Section II.A.2.c.

174. *See* Tim Hwang, *Dealing with Disinformation: Evaluating the Case for CDA 230 Amendment Interventions,* Stanford PACS 32, https://pacscenter.stanford.edu/dealing-with-disinformation-evaluating-the-case-for-cda-230-amendment-interventions/ (last visited Mar. 31, 2021) [https://perma.cc/5TL5-J4S8].

175. *See* Chesney & Citron, *supra* note 7, at 1798–99.

176. *See id.*

177. *See, e.g.*, Matthew Ferraro, *Congress's Deepening Interest in Deepfakes*, THE HILL (Dec. 29, 2020, 12:00PM), https://thehill.com/opinion/cybersecurity/531911-congresss-deepening-interest-in-deepfakes [https://perma.cc/W9R6-3ALF] (the Identifying Outputs of Generative Adversarial Networks (IOGAN) Act requires the Director of the National Science Foundation to support research and the National Institute of Standards and Technology to develop standards for examining deepfakes).

178. *See id.* (Virginia criminalized the distribution of nonconsensual deepfake pornography in 2019, establishing a maximum one year in jail and $2,500 fine).

On December 20, 2019, the National Defense Authorization Act (NDAA) for the 2020 fiscal year weighed in on the deepfake debate.[179] The NDAA requires the Director of National Intelligence (DNI) to submit a comprehensive report on the foreign weaponization of deepfakes to Congressional Intelligence Committees.[180] The DNI must also notify Congress of foreign deepfake disinformation activities specifically targeting the U.S. election process.[181] The DNI is also authorized to award up to $5 million to encourage development of deepfake detection technology.[182]

Deepfakes are not just a problem in the U.S., and other countries have adopted their own legislation to tackle the mounting challenges deepfakes present. Deepfakes have been prevalent in China, for example, a country that might consume as much information as the U.S. China recently banned online video and audio providers from using deepfakes, citing concerns over the growing disinformation war occurring globally.[183] The ban further extends to both providers and users of online video news and audio services from using or distributing deepfakes or fake news.[184] Providers and users of online video news and audio information services must label any content that involves new technologies such as deep learning.[185] Content providers must also use technology to detect manufactured or manipulated content in violation of the regulation.[186] China's ban encompasses deepfakes used in the political sense and any other area deepfakes might emerge, such as virtual reality.[187] China's deepfake ban appears to ban deepfakes writ large, even creative or artistic ones, and includes consequences for refusal to comply.

While the U.S. would not likely enact laws similar to those of China, it is helpful to see another country's approach to the rising problem of deepfakes. The U.S. is presented with its own challenges in combatting deepfakes, but the legislation currently enacted is a step in the right direction. A stronger approach will be needed in the coming years, but scientists and technologists are trying to come up with their own solution in the meantime.

C. Fighting Technology with Technology

Algorithms and artificial intelligence might seem like an attractive solution to moderating content online at first blush, but there is a plethora of issues that arise when AI is involved.[188] Unfortunately, we are not at the point

179. *Id.*
180. *Id.*
181. *Id.*
182. Ferraro, *supra* note 177.
183. Meng Jing, *China Issues New Rules to Clamp Down on Deepfake Technologies Used to Create and Broadcast Fake News,* S. China Morning Post (Nov. 29, 2019), https://www.scmp.com/tech/apps-social/article/3039978/china-issues-new-rules-clamp-down-deepfake-technologies-used [https://perma.cc/77ZC-TMWR] (providing that China has also voiced concerns over deepfakes from creation of the face-swap app Zao).
184. *Id.*
185. *Id.*
186. *Id.*
187. *Id.*
188. *See* Chesney & Citron, *supra* note 7, at 1787.

yet where AI returns highly accurate takedown responses.[189] In the event that AI makes a mistake, it runs the risk of violating free speech by filtering out protected speech and media.

Deepfake scanners and other video editing software might be a more attractive approach.[190] Researchers are starting to create tools that attempt to dissect deepfake videos and distinguish the real from the fake. For example, Binghamton University in New York has teamed up with Intel to create "FakeCatcher," a tool that reveals deepfakes by discovering subtle differences in skin color caused by the human heartbeat.[191] Social media companies should implement such deepfake detection software on their sites. Users should also be able to challenge the software's finding of a deepfake if they believe it was in error.

Another moderating option is blockchain, a popular resource for authenticating business and financial records. Blockchain can be used for authenticating videos.[192] Using blockchain technology, and when a video is uploaded to a site, the metadata from the video would be captured (including the upload time, location, and creator/uploader's ID), which would create a transparent and traceable route proving the authenticity of the video.[193] Any fake, copy, or change to the video would be noted through the blockchain by that video's own unique metadata.[194] The technology is out there. Social media companies just have to engage with the researchers developing it to combat the manipulative media together.

D. Knowledge is Power

While we might not be able to stop the oncoming threat of deepfakes, we can at least start implementing the tools to help increase awareness of deepfakes. The problem with deepfakes is that they reflect a bigger problem within society itself, stemming from a general lack of trust in public officials and our basic democratic institutions.[195] But because we are already faced with similar problems like fake news, deepfakes might be the wake-up call we need to help fix disinformation in our society.[196]

Deepfakes are gaining prominence as creative, innovative tools, but not all consumers know about them. If more people become aware of the

189. *See* Mark Scott & Laura Kayali, *What Happened When Humans Stopped Managing Social Media Content*, POLITICO (Oct. 21, 2020), https://www.politico.eu/article/facebook-content-moderation-automation/ [https://perma.cc/UNR6-YUY9].

190. *See, e.g.*, *About Us*, DEEPWARE, https://deepware.ai/about/ (last visited Mar. 31, 2021) [https://perma.cc/Q27P-PBJ6].

191. *See* Chris Kocher, *Best Way to Detect 'Deepfake' Videos? Check for the Pulse*, BINGHAMTON UNIV. (Oct. 21, 2020), https://www.binghamton.edu/news/story/2713/best-way-to-detect-deepfake-videos-check-for-the-pulse [https://perma.cc/9HKW-X7BN].

192. *See* Jason Tashea, *Some States Are Allowing People and Companies to Use Blockchain to Authenticate Documents*, ABA J. (Sept. 1, 2019), https://www.abajournal.com/magazine/article/best-evidence [https://perma.cc/GF4L-AHTR].

193. *Id.*

194. *Id.*

195. *See* Silbey & Hartzog, *supra* note 6, at 964.

196. *Id.*

existence of deepfakes, they will be less likely to be fooled by one. Learning how to evaluate facts, test systems, and challenge accounts by examining alternative perspectives is a way to turn people into deep thinkers, and in turn deep thinkers will not be so easily fooled by deepfakes.

V. CONCLUSION

Deepfakes arguably have creative expressive values, and if we learn how to filter out the harmful deepfakes from the harmless, society will benefit. Current legal remedies are inadequate because of the timing and nature of deepfakes. Deepfakes are most damaging when people are exposed to them and believe their lies. However, an outright ban on deepfakes is impossible in light of the First Amendment. But amid all of these challenges, social media companies are working with academics, the government, and other leaders in the technology industry to create adoptable solutions, and they should continue to do so. Other remedies, such as amending Section 230 or state laws are likewise feasible. Although the perfect solution is not here yet, it is in sight. I will believe it when I see it.