HAVE WE NO DECENCY? SECTION 230 AND THE LIABILITY OF SOCIAL MEDIA COMPANIES FOR DEEPFAKE VIDEOS

NICHOLAS O'DONNELL*

Deepfake videos, made using machine learning technology, have the potential to increase the efficacy of disinformation campaigns. Deepfakes can be broadly classified into two categories. The first category, which includes Deepfake pornography, has a harmful effect regardless of the number of individuals who view the videos. The second category, and the one addressed by this Note, requires broad dissemination to have a harmful impact. Three specific Deepfake threats are analyzed: (1) election interference; (2) economic interference; and (3)public safety. Existing legal mechanisms are insufficient to address the threat because Section 230 of the Communication Decency Act shields social media companies from liability for Deepfakes disseminated on their platforms. Even proposed amendments to Section 230 do not adequately address the Deepfake threat. This Note will argue that Deepfake videos are not protected speech under the First Amendment, and social media companies can be considered publishers of Deepfakes if the videos are not quickly removed. Section 230 should thus be amended to create a narrow exception for Deepfake videos. Regulatory agencies should then enact new rules and amend existing ones in order to hold social media companies liable for the circulation of Deepfakes. The threat of liability will deter social media companies from allowing the videos to spread unchecked on their platforms and incentivize them to develop new technology for prompt detection and removal.

TABLE OF CONTENTS

[*] I would like to thank Professor Jason Mazzone for advising me as I wrote this Note. He challenged me to find creative solutions to the problems I encountered in my early drafts, and his insights and feedback were invaluable.

I. INTRODUCTION

At 1:00 AM, the morning of the 2020 presidential election, a video is uploaded by a largely unknown Twitter user, with the title "Campaign Manager Announces Democratic Candidate's Death." Skeptical users, allured by the title, click play only to see exactly what the title described: the candidate's campaign manager solemnly announcing that the previous evening the candidate died of a heart attack. The few followers awake to view the video share it, and it quickly goes viral. By 5:00 AM, the video has been shared over 100,000 times and viewed by several million individuals from all over the world. It has spread across every major social media platform. While every media outlet reports that the video is a fake, the many Americans who get their news primarily from social media[1] plan their days believing that a major party does not have a candidate in the general election. The effects are compounded by several cites refusing to remove the video or even disclaim that it is fake.[2] Turnout is similar to the 2016 election,[3] but with surprising differences: turnout in swing states heavily targeted by both parties was substantially smaller than expected, and third-party candidates received the highest aggregate number of votes in modern American history. By midnight, two major stories dominate the news: Donald Trump has won

[1] *See* Elisa Shearer, *Social Media Outpaces Print Newspapers in the U.S. as a News Source*, PEW RSCH. CTR. (Dec. 10, 2018), https://www.pewresearch.org/fact-tank/2018/12/10/social-media-outpaces-print-newspapers-in-the-u-s-as-a-news-source/ [https://perma.cc/EQ95-UTT3].

[2] When a "shallowfake" of Pelosi was widely shared in 2019, YouTube removed the video, Facebook merely directed users to reports calling it a fake, and Twitter did nothing. Makenna Kelly, *Congress Grapples with How to Regulate Deepfakes*, VERGE (June 13, 2019, 1:30 PM), https://www.theverge.com/2019/6/13/18677847/deep-fakes-regulation-facebook-adam-schiff-congress-artificial-intelligence [https://perma.cc/YE75-VGVN].

[3] Thom File, *Voting in America: A Look at the 2016 Presidential Election,* CENSUS.GOV: RANDOM SAMPLINGS (May 10, 2017), https://www.census.gov/newsroom/blogs/random-samplings/2017/05/voting_in_america.html [https://perma.cc/KM52-UYXW].

re-election, and a Russian troll farm successfully deployed deepfake technology to interfere with the American election.[4]

Deepfake videos are produced from various video-editing technologies that allow users to create fake, realistic videos of individuals from a cache of photographic images through the use of machine learning and neural networks.[5] Current technology is still in its infancy,[6] but it is quickly becoming more sophisticated and convincing.[7] Despite the ubiquitous discussion of the threat of "fake news,"[8] deepfake videos seem to be a substantively different problem. Even if news stories deploy fake facts or statements, video remains a bastion of veracity.[9] Regardless of what someone else *says*, the ability to *see* something for one's self is uniquely convincing.[10]

Generally, the threats posed by deepfake videos can be broken into two categories. The first category contains those videos whose impact only requires a limited number of people to see it or share it.[11] These include, for example, deepfake pornography where the mere existence of a video is an invasion of privacy and dignity.[12] Similarly, deepfake security intelligence can undermine the operations of the military and intelligence agencies even if they are only viewed by a few individuals.[13]

The second category, in contrast, requires a deepfake video to penetrate public discourse.[14] These impacts require a large number of people to view a video and believe it to be true.[15] A video posted on the eve of an election, for

[4] This "election-eve" scenario is based on the testimony of professor Danielle Citron's testimony before the House Select Committee on Intelligence. *The National Security Challenge of Artificial Intelligence, Manipulated Media, and 'Deepfakes': Hearing Before the H. Permanent Select Comm. on Intel.*, 116th Cong. 4 (2019), https://docs.house.gov/Committee/Calendar/ByEvent.aspx?EventID=109620 [https://perma.cc/V86W-EE67] (prepared written testimony and statement of Danielle K. Citron, Morton & Sophia Macht Professor of Law, University of Maryland Carey School of Law). While the scenario in this case involves the Democratic candidate, it is just as easy to imagine a video circulating that shows FBI Director Christopher Wray announcing a criminal investigation targeting Trump for tax fraud.

[5] Bobby Chesney & Danielle K. Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1759–60 (2019); Regina Rini, *Deepfakes and the Epistemic Backstop*, 20 PHILOSOPHERS' IMPRINT 1 (2020).

[6] *See* Rini, *supra* note 5, at 1.

[7] *Id.* at 6–7; Nina I. Brown, *Deepfakes and the Weaponization of Disinformation*, 23 VA. J. L. & TECH. 1, 5–6 (2020).

[8] *See generally*, Xichen Zhang & Ali A. Ghorbani, *An Overview of Online Fake News: Characterization, Detection, and Discussion*, 57 INFO. PROCESSING & MGMT. 1 (2020) (discussing fake news and the origins of fake news discourse following the 2016 election).

[9] Robert Chesney & Danielle Keats Citron, *21st Century Style Truth Decay: Deep Fakes and the Challenge for Privacy, Free Expression, and National Security*, 78 MD. L. REV. 882, 883–84 (2019).

[10] *Id.*

[11] *See*, *e.g.*, Mary Anne Franks & Ari Ezra Waldman, *Sex Lies and Videotape: Deep Fakes and Free Speech Delusions*, 78 MD L. REV., 892, 893 (2019) (discussing the inherent harm of deepfake pornography).

[12] *See id.*; Danielle Keats Citron & Mary Anne Franks, *Criminalizing Revenge Porn*, 49 WAKE FOREST L. REV. 345, 346 ("[N]onconsensual pornography . . . [is] an egregious privacy violation that deserves criminal punishment.").

[13] *See* Chesney & Citron, *supra* note 5, at 1783–84 (discussing the use of deepfakes for strategic, operational, and tactical deception against military and intelligence agencies).

[14] Brown, *supra*, note 7, at 14–15.

[15] *Id.*

example, will not have much of an impact if it is only viewed by several hundred people across the country. While election interference is the most commonly discussed scenario,[16] it is also possible that a well-placed deepfake could disrupt an initial public offering,[17] or cause massive civil unrest[18] and undermine law enforcement.[19]

This Note addresses the second category of threats. While many commentators and politicians recognize the potential harms of deepfake videos, the federal government has largely ceded action on this matter to the states and to companies themselves.[20] A piece of legislation proposed in 2018 has no momentum,[21] and a June 2019 hearing by the House Select Committee on Intelligence yielded a bill unlikely to become law.[22] The limitations of federal solutions are due, in part, to Section 230 of the Communication Decency Act which provides immunity to social media companies for content posted on their sites.[23] If social media companies cannot be liable for the content, any solution must focus on the creators and distributors of deepfake videos.[24] Even if Section 230 were to be amended, any regulation of deepfakes must comply with the First Amendment.[25] Deepfakes, as a kind of video-editing technology, are also a kind of expression,[26] so a poorly crafted regulation would be unconstitutional.[27] This Note argues Section 230 should be amended to treat online platforms as the publishers of deepfake videos posted to their sites. This would allow federal regulatory agencies to issue fines to social media companies for deepfakes distributed on their platforms.[28] The possibility of large federal fines has several benefits

[16] *See*, *e.g.*, Chesney & Citron, *supra* note 5, at 1778–79.

[17] Citron, *supra* note 4, at 5.

[18] Chesney & Citron, *supra* note 5, at 1780–81.

[19] *Id.* at 1779.

[20] *See* Levi Sumagaysay, *California Has a New Deepfakes Law in Time for 2020 Election*, MERCURY NEWS (Oct. 4, 2019, 11:36 AM), https://www.mercurynews.com/2019/10/04/california-has-a-new-deepfakes-law-in-time-for-2020-election/ [https://perma.cc/4PSK-TQF9] (discussing laws in California and Texas). For a proposed New York law, see S. 5959D 2019—2020 Reg. Sess. (N.Y. 2019).

[21] *See* Malicious Deep Fake Prohibition Act of 2018, S. 3805, 115th Cong. (Dec. 21, 2018).

[22] For the law, see Deepfakes Accountability Act, H.R. 3230, 116th Cong. (June 12, 2019). For an overview of the hearing, see Emily Tillet & Olivia Gazis, *House Holds Hearing on "Deepfakes" and Artificial Intelligence Amid National Security Concerns*, CBS NEWS (June 13, 2019, 11:16AM), https://www.cbsnews.com/news/house-holds-hearing-on-deepfakes-and-artificial-intelligence-amid-national-security-concerns-live-stream/ [https://perma.cc/BGJ4-BBHN]. Full video of the hearing is available at *Open Hearing on Deepfakes and Artificial Intelligence*, YOUTUBE (June 13, 2019), https://www.youtube.com/watch?v=tdLS9MlIWOk [https://perma.cc/9QEW-WXJB].

[23] 47 U.S.C. § 230.

[24] *See* Cristiano Lima, *"Nightmarish": Lawmakers Brace for Swarm of 2020 Deepfakes*, POLITICO (June 13, 2019, 5:04 AM), https://www.politico.com/story/2019/06/13/facebook-deep-fakes-2020-1527268 [https://perma.cc/P4A5-AJU9] (quoting Adam Schiff discussing the utility of amending section 230).

[25] Rebecca Green, *Counterfeit Campaign Speech*, 70 HASTINGS L.J. 1445, 1448–49 (2019).

[26] Rebecca A. Delfino, *Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act*, 88 FORDHAM L. REV. 887, 925 (2019).

[27] Green, *supra* note 25, at 1448–49.

[28] *See* discussion *infra* Section IV.

over a civil or criminal system of liability and will incentivize companies to remove deepfakes before they can be circulated or to remove them after they have been posted.[29] Such a system would also survive a constitutional challenge.[30]

Part II of this Note will explain deepfake technology and elaborate on the specific threats that it poses. It will then discuss proposed solutions to these problems, from existing legislation to proposed criminal and civil penalties. Additionally, the immunity granted to social media companies under Section 230 of the Communications Decency Act ("CDA") will be considered. Part III will then analyze the limitations of these solutions. It will discuss their failings at the policy level, as well as the constitutional barriers any regulation of deepfake videos will face. Malicious deepfake videos will be shown to be unprotected speech, and the possibility of a regulatory agency suing online platforms will be considered.

Finally, Part IV will outline a specific amendment to Section 230 and show how it will incentivize social media companies to remove deepfake videos before they can be widely circulated. Section 230 should be amended so that online platforms are treated as the publishers of deepfake videos posted on their sites. This would allow federal regulatory agencies to issue fines based on the kind of deepfake that is distributed. Various governmental agencies would be considered as enforcement agents. Social media companies are in the best position to prevent the spread of deepfake videos and the possibility of steep fines will ensure actors comply with new regulations. Different techniques to police deepfake videos, from machine learning technology to mechanisms borrowed from copyright law, will be analyzed for strengths and weaknesses. Ultimately, extending liability to social media companies for the impacts of deepfake videos will incentivize the companies to institute a variety of mechanisms that will successfully prevent deepfake videos from penetrating mainstream discourse.

II. BACKGROUND: THE EVOLUTION OF DEEPFAKE TECHNOLOGY, ITS THREATS, AND EXISTING SOLUTIONS

Deepfakes are fabricated video or audio recordings created through machine learning technology.[31] A computer program uses a large dataset of real audio-visual recordings to build a model of the facial and vocal characteristics of a person, then it superimposes this model onto recordings of another person.[32] The effect is an apparent recording of a well-known person doing or saying something that never occurred.[33] While an exact definition of what constitutes a deepfake video is difficult to give, this Note adopts the one given by professors Robert Chesney and Danielle Citron, which states:

[29] *See infra* Section III.A (discussing the limits of criminal and civil liability); Section IV (discussing how social media companies will react to regulatory liability).

[30] *See* discussion *infra* Section III.C.

[31] Franks & Waldman, *supra* note 11, at 893–95.

[32] *Id*.

[33] Chesney & Citron *supra* note 5, at 1758; Rini, *supra* note 5, at 1.

> [P]ornographers have been early adopters of the relevant technology, interposing the faces of celebrities into sex videos. This has given rise to the label 'deep fake' for such digitized impersonations. We use that label here more broadly, as a shorthand for the full range of hyper-realistic digital falsification of images, video, and audio . . . . Deep-fake technology . . . leverages machine-learning algorithms to insert faces and voices into video and audio recordings of actual people and enables the creation of realistic impersonations out of digital whole-cloth. The end result is realistic-looking video or audio making it appear that someone said or did something.[34]

Before analyzing the shortcomings of existing and proposed solutions to the spread of deepfake videos, it is necessary to examine the technology's history, its specific political and economic threats, and the policy mechanisms that have been previously advocated.

A. From Reddit to Star Wars: A Brief History of Deepfake Videos

Deepfake technology first emerged in a Reddit thread that contained fake pornographic videos of celebrities; the original poster of the videos called themselves "deepfakes."[35] According to deepfakes, the videos were created using open source machine learning technology, like Keras and TensorFlow.[36] The videos relied on a cache of images taken from Google image search, stock photos, and YouTube videos.[37] The software utilized deep learning technology that ran computations on input data taken from both celebrities' faces and the pornographic videos onto which they were superimposed.[38] After some "training," the algorithm manipulating the videos was able to finish the manipulation without human guidance.[39]

While the videos were not entirely convincing[40] and the thread was ultimately taken down by Reddit,[41] deepfakes's thread was a watershed moment in that it alerted people to the possibility of completely fabricated videos being widely circulated on the internet.[42] Concern about the spread of the technology has proven justified as deepfakes have become increasingly popular: between December 2018 and July 2019, the number of deepfake videos on the internet

[34] Chesney & Citron, *supra* note 5, at 1757–58.

[35] Benjamin Goggin, *From Porn to "Game of Thrones": How Deepfakes and Realistic-Looking Fake Videos Hit It Big*, BUS. INSIDER (Jun. 23, 2019, 9:45 AM), https://www.businessinsider.com/deepfakes-explained-the-rise-of-fake-realistic-videos-online-2019-6 [https://perma.cc/MJT5-EGCC]; Emma Grey Ellis, *People Can Put Your Face on Porn—And the Law Can't Help You*, WIRED (Jan. 26, 2018, 7:00 AM), https://www.wired.com/story/face-swap-porn-legal-limbo/ [https://perma.cc/RJ6J-Q2BH].

[36] Goggin, *supra* note 35.

[37] *See id.*

[38] *See id.*

[39] *See id.*

[40] *See id.*

[41] Jaime Dunaway, *Reddit (Finally) Bans Deepfake Communities, but Face-Swapping Porn Isn't Going Anywhere*, SLATE (Feb. 8, 2018, 4:27 PM), https://slate.com/technology/2018/02/reddit-finally-bans-deepfake-communities-but-face-swapping-porn-isnt-going-anywhere.html [https://perma.cc/3FZN-EY4S].

[42] *See generally* Ellis, *supra* note 35 (discussing the implications of technology like that used in the deepfakes thread, as well as its probable persistence into the future).

doubled.[43] Although many versions of the technology are still in their infancy, they are improving[44] and are beginning to have a real-world impact in contexts beyond fake pornography.[45] While not an entirely fabricated video, an edited "shallowfake" of Nancy Pelosi, in which she appeared to be drunk, was widely shared in 2019, even being retweeted by President Trump and his lawyer Rudy Giuliani.[46] A fake Trump pee-tape, which was possibly created using deepfake technology, has been circulating online since at least January 2019.[47] The Flemish socialist party posted a deepfake video in which Trump urged Belgium to withdraw from an environmental treaty.[48] The most sophisticated use of this technology, however, can be seen in *Star Wars: Rogue One*.[49] In a scene at the end of the movie, intercepted plans for the Death Star are delivered to Princess Leia, who bears the likeness of Carrie Fisher.[50] But Fisher never acted in this scene: when Leia is seen from the front, the viewer is looking at a purely digital recreation, and, when she is viewed from behind, Leia is played by an entirely different actor, Ingvild Deila.[51]

Given the existing uses of deepfake technology, its potential threats become apparent. If a malicious actor deployed it using sophisticated means, or if a significant breakthrough were made in open source software,[52] a video could be circulated that would be undetectable as a deepfake to the average viewer. A well placed deepfake video could exacerbate already-exiting problems caused by disinformation to a crisis point. Specifically, the intrusion of a sophisticated deepfake video into the mainstream could threaten election integrity, economic stability, and public safety.

[43] *The State of Deepfakes: Landscape, Threats, and Impact*, SENSITY 1, 1 (Sept. 2019), https://medium.com/sensity/tracer-newsletter-31-07-10-19-the-state-of-deepfakes-landscape-threats-and-impact-9cdee0bc13e6 [https://perma.cc/YPG3-JL6M].

[44] Chesney & Citron, *supra* note 5, at 1759–61 (discussing various technologies used in deepfake videos and their advancements); Rini, *supra* note 5, at 11–12.

[45] *See generally* SENSITY, *supra* note 43 (discussing the impact of several deepfake videos).

[46] *Id.* at 11; Drew Harwell, *Faked Pelosi Videos, Slowed to Make Her Appear Drunk, Spread Across Social Media*, WASH. POST (May 24, 2019, 3:41 PM), https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/?utm_term=.5bd6ddefb248 [https://perma.cc/Z8HT-8ZW2]. A "shallowfake" "refers to videos that have been manipulated with basic editing tools or intentionally placed out of context." SENSITY, *supra* note 43, at 11.

[47] Dan Robitzski, *Yes, There's a "Pee Tape"—and It's Unclear If It's a Deepfake*, FUTURISM (Sept. 26, 2019), https://futurism.com/the-byte/pee-tape-unclear-deepfake [https://perma.cc/S7NA-SAR8]; Ashley Feinberg, *The Pee Tape Is Real, But It's Fake*, SLATE (Sept. 25, 2019, 9:11 PM), https://slate.com/news-and-politics/2019/09/inside-the-convincing-fake-trump-pee-tape.html [https://perma.cc/L2VP-Y2GV].

[48] Rini, *supra* note 5, at 5.

[49] Dave Itzkoff, *How Rogue One Brought Back Familiar Faces*, N.Y. TIMES (Dec. 27, 2016), https://www.nytimes.com/2016/12/27/movies/how-rogue-one-brought-back-grand-moff-tarkin.html [https://perma.cc/63GZ-NGG2].

[50] *Id.*

[51] *Id.*

[52] Chesney & Citron, *supra* note 5, at 1757–58.

B. The Threat of Deepfakes: Election Interference, Economic Disinformation, and Civil Unrest

The most discussed threat of deepfake videos is the possibility of election interference. In June 2019, the House Intelligence Committee held a hearing to discuss the threat of deepfake videos and artificial intelligence, which emphasized the potential of the technology to disrupt an election.[53] Both state and nonstate actors could potentially use deepfakes to interfere with American elections. Regarding state actors, Russia could use deepfake videos to expand its existing election interference tools.[54] Russia has previously used fake documents when trying to influence elections—during the 2017 French election, Russia stole documents from the Macron campaign and edited them to include fake, damaging information.[55] While the effort failed to have an effect (due, in part, to Russian incompetence), it is easy to imagine the use of a deepfake in a broader disinformation campaign.[56] If a deepfake were deployed along with fabricated news articles, it would, presumably, increase the efficacy of any interference.[57] Besides Russia, any state actor that wanted to meddle in American elections, such as China or Iran, could use deepfakes as part of their interference campaign.[58]

Nonstate actors could also harness deepfakes to interfere with elections. Oligarchs, corporations, and activists could marshal deepfakes for their desired end.[59] Given that the threat of interference comes from the public mistaking a deepfake video for reality, a private actor could sway public opinion as successfully as a State.[60] The motivation certainly exists, and the only existing limitations are financial and technological.[61] If these groups received greater financial

[53] Tillet & Gazis, *supra* note 22 ("Committee chair Rep. Adam Schiff said the spread of manipulated videos presents a 'nightmarish' scenario for the 2020 presidential elections"); Kelly, *supra* note 2.

[54] Chesney & Citron, *supra* note 5, at 1778–79; Citron, *supra* note 4, at 5; 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 [https://perma.cc/7J6Q-7EDG].

[55] Chesney & Citron, *supra* note 5, at 1778.

[56] *Id.* at 1779.

[57] Brown, *supra* note 7, at 7–8.

[58] *The National Security Challenges of Artificial Intelligence, Manipulated Media, and 'Deepfakes': Hearing Before the H. Permanent Select Comm. on Intel.*, 116th Cong. 1–2 (2019) (statement of Clint Watts, Distinguished Research Fellow, Foreign Policy Research Institute), https://docs.house.gov/meetings/IG/IG00/20190613/109620/HHRG-116-IG00-Wstate-WattsC-20190613.pdf [https://perma.cc/5ZMY-KDU9] (statement of Clint Watts, Distinguished Research Fellow, Foreign Policy Research Institute); Khari Johnson, *Deepfake Concerns Ahead of 2020 Election Include Iran, China, Instagram, and WhatsApp*, VENTURE BEAT (Sept. 3, 2019, 12:30 PM), https://venturebeat.com/2019/09/03/deepfake-concerns-ahead-of-2020-election-include-iran-china-instagram-and-whatsapp/ [https://perma.cc/8AA6-SZGU].

[59] Watts, *supra* note 58, at 2; Chesney & Citron, supra note 5, at 1779.

[60] *See National Security Challenges of Artificial Intelligence, Manipulated Media, and "Deepfakes": Hearing Before the H. Permanent Select Comm. on Intel*, *supra* note 58, at 2; *see also* Clint Watts, *Advanced Persistent Manipulators, Part One: The Threat to the Social Media Industry*, ALL. FOR SECURING DEMOCRACY (Feb. 12, 2019) [hereinafter "Watts, *Advanced Persistent Manipulators*"], https://securingdemocracy.gmfus.org/advanced-persistent-manipulators-part-one-the-threat-to-the-social-media-industry/ [https://perma.cc/FB9R-MEWG] (discussing persistent efforts by extremist and activist groups to use social media as a part of disinformation campaigns).

[61] Chesney & Citron, *supra* note 5, at 1779; *Watts, Advanced Persistent Manipulators*, *supra* note 60.

backing, or if a technological breakthrough made sophisticated video manipulation technology readily accessible, then nonstate actors could readily diffuse a convincing deepfake video.[62]

Additionally, deepfake videos could cause significant economic disruption. Similar to election meddling scenarios, a deepfake could be released the night before a company's initial public offering, causing the stock price to fall with tens, if not hundreds, of millions of dollars lost.[63] Additionally, sophisticated deepfake technology would allow hostile actors to create videos of directors and executives causing creditors and investors to make rash decisions.[64] After a video of Elon Musk smoking marijuana on a podcast became public, Tesla's stock price fell 6%.[65] While Musk actually engaged in the recorded behavior, deepfake technology would allow individuals to make similar videos in which "puppets" of corporate directors and executives acted outlandishly.[66]

Certain actors are already deploying fake economic information to manipulate markets. "Disinformation-for-hire" companies advertise on underground Russian forums and offer potential clients a variety of services, from commenting on articles to posting on social media.[67] These companies are beginning to target Western businesses.[68] Some of these companies advertise their ability to place fake stories in major media outlets including Reuters and Mashable.[69] From this background, deepfake-for-hire services are not difficult to imagine. Tesla has already been the target of a fake video: in January 2019, a video was widely circulated on the internet showing a self-driving car crashing into a robot prototype at a consumer electronics show.[70] The video was staged and may have been created as part of a promotional strategy or a disinformation campaign.[71] Regardless of its origin, the video is a cautionary testament to the ability for deepfakes to damage a company's reputation, and a future video may have severe economic consequences.

Finally, deepfakes could pose a significant threat to public safety by provoking civil unrest. Deepfake videos could be used to manufacture crises or to inflame tensions during a protest. The Russian Internet Research Agency has launched sophisticated campaigns to create the appearance of a chemical disaster

[62] Chesney & Citron, *supra* note 5, at 1779; Watts, *supra* note 58, at 2.

[63] Citron, *supra* note 4, at 5.

[64] *See* Henry Ajder, *Social Engineering and Sabotage: Why Deepfakes Pose an Unprecedented Threat to Business*, SENSITY (Mar. 10, 2019),), https://medium.com/sensity/scams-and-sabotage-why-deepfakes-pose-an-unprecedented-threat-to-businesses-537875524b31 [https://perma.cc/S8VS-9E5F].

[65] *Id.*

[66] *Id.*

[67] Ben Popken, *Trolls for Hire: Russia's Freelance Disinformation Firms Offer Propaganda with a Professional Touch*, NBC NEWS (Oct. 1, 2019, 10:40AM), https://www.nbcnews.com/tech/security/trolls-hire-russia-s-freelance-disinformation-firms-offer-propaganda-professional-n1060781 [https://perma.cc/3D2G-DJH2].

[68] *Id.*

[69] Insikt Grp., *The Price of Influence: Disinformation in the Private Sector*, RECORDED FUTURE 1, 11 (2019), https://www.recordedfuture.com/disinformation-service-campaigns/ [https://perma.cc/CME6-55ST].

[70] Claire Atkinson, *Fake News Can Cause "Irreversible Damage" to Companies– and Sink Their Stock Price*, NBC NEWS (Apr. 25, 2019, 11:54 AM), https://www.nbcnews.com/business/business-news/fake-news-can-cause-irreversible-damage-companies-sink-their-stock-n995436 [https://perma.cc/5254-YLT4].

[71] *Id.*

in Louisiana and an Ebola outbreak in Atlanta.[72] Such an attempt to sow disinformation could be compounded by, for example, a deepfake video showing the head of the Center for Disease Control describing the outbreak in painstaking detail.[73]

This sort of disinformation has the capacity to influence a staggering number of individuals. During the "yellow-jacket" protests in France, disinformation about the protests was viewed nearly 100 million times on Facebook, with much of it coming from the Russian television network RT.[74] RT videos of the protests were viewed twenty-three million times, more than all mainstream French news outlets combined.[75] While the impacts of the disinformation was small—only affecting, at most, tourism[76]—had a deepfake been targeted at and circulated by the protestors themselves, the effects could have been disastrous, possibly leading to violent clashes with police.

If such a video were successfully introduced in America, the effects are readily apparent.[77] A video could be introduced showing Immigrations and Customs Enforcement agents mistreating a child, or FBI agents talking about how to remove Trump from office.[78] At best, these videos would inflame existing social tensions, while at worst they would cause mass demonstrations to erupt into violence. Russia successfully organized a fake protest prior to the 2016 election that was attended by thousands of people in New York,[79] and another in Florida.[80] The only apparent purpose of these protests was to deepen existing tensions,[81] believing that those divisions would prevent consensus about policy.[82] Deepfakes would strengthen these already-existing efforts.[83] Deepfakes may not only be used in organizing demonstrations, they could be used to escalate tensions in fraught situations. If a deepfake were used to inflame tensions during a protest, it is very possible a peaceful protest could turn into a riot.[84]

[72] Chesney & Citron, *supra* note 5, at 1782.

[73] *Id.*

[74] Atkinson, *supra* note 70.

[75] *Id.*

[76] *Id.*

[77] Chesney & Citron, *supra* note 5, at 1779.

[78] *Id.*

[79] Ali Breland, *Thousands Attended Protest Organized by Russians on Facebook*, HILL (Oct. 31, 2017, 1:15 PM), https://thehill.com/policy/technology/358025-thousands-attended-protest-organized-by-russians-on-facebook [https://perma.cc/3TSK-VUFF].

[80] BLOOMBERG, *Russians Staged Rallies for and Against Trump to Promote Discord, Indictment Says*, FORTUNE (Feb. 17, 2018, 11:55 AM), https://fortune.com/2018/02/17/russian-organized-rallies-election-meddling/ [https://perma.cc/9W7G-33C8].

[81] *See id.*

[82] Chesney & Citron, *supra* note 5, at 1780.

[83] *Id.*

[84] *See* Brown, *supra* note 7, at 7, 11.

C. A Patchwork Approach–The Myriad Mechanisms for Controlling the Spread of Deepfakes

While there is a growing consensus about the threat of deepfake videos,[85] the solution is not easily identifiable. Some existing and proposed solutions attempt to extend civil and criminal liability to deepfake creators without amending Section 230.[86] These solutions have appeared in both enacted and pending legislation.[87] Some commentators have proposed extending existing civil liability categories so that creators of deepfakes could be sued, while others have argued that Section 230 should be amended to extend civil liability to social media companies.[88]

No existing federal law explicitly targets the use of deepfake technology, but two bills have been introduced in the United States Senate and House.[89] Senator Ben Sasse proposed the Malicious Deepfakes Prohibition Act in 2018, which would broadly criminalize the creation and distribution of malicious deepfakes,[90] but it has since expired.[91] Conversely, a bill currently in the House would require watermarking and labelling of all deepfake videos and would allow victims to seek civil penalties and an injunction against the videos' creators.[92] Other proposed federal legislation, while not explicitly targeting deepfakes, would support research into them[93] and require briefing and support to Congress.[94]

While no federal law has been enacted, several states have successfully passed legislation targeting deepfakes in some capacity. California has passed a law that would prohibit deepfakes made "with the intent to injure [a] candidate's reputation or to deceive a voter into voting for or against the candidate," unless

[85] *See* Chesney & Citron *supra* note 5, at 1757 n.5.

[86] *See infra* notes 102–108 and accompanying text; *see generally* Green, *supra* note 25 (discussing the possibility of criminalizing false campaign speech).

[87] *See infra* notes 89–101 and accompanying text.

[88] *See infra* notes 120–126 and accompanying texts.

[89] *See* Malicious Deep Fake Prohibition Act of 2018, S. 3805, 115th Cong. (Dec. 21, 2018); DEEP FAKES Accountability Act, H.R. 3230, 116th Congress (June 12, 2019) (referred to Subcommittee on Crime, Terrorism, and Homeland Security).

[90] S. 3805; *see also* Kaveh Waddell, *Lawmakers Plunge into "Deepfake" War*, AXIOS (Jan. 31, 2019), https://www.axios.com/deepfake-laws-fb5de200-1bfe-4aaf-9c93-19c0ba16d744.html [https://perma.cc/A6KQ-99SZ].

[91] Matthew F. Ferraro, *Deepfake Legislation: A Nationwide Survey—State and Federal Regulators Consider Legislation to Regulate Manipulated Media*, WILMER HALE 1, 6–7 (Sept. 25, 2019). Sasse has also introduced other bills on the topic of deepfakes, but they are not as broad and deal specifically with members of the armed forces, federal employees, and their families. *See id.*

[92] DEEP FAKES Accountability Act, .R. 3230, 116th Cong. (2019); Ferraro, *supra* note 91, at 7–9 (discussing the civil penalties and injunctions, as well as other reporting requirements); *see also* Nina Iacono Brown, *Congress Wants to Solve Deepfakes by 2020*, SLATE (July 15, 2019, 7:30 AM), https://slate.com/technology/2019/07/congress-deepfake-regulation-230-2020.html [https://perma.cc/6LML-4HUS].

[93] *See* Ferraro, *supra* note 91, at 3.

[94] *Id.* at 4–6. While not sent in support of a particular piece of legislation, a letter from Senators Marco Rubio and Mark Warner was sent to eleven social media companies, emphasizing the deepfakes threat and asking the companies what measures they were taking to combat it. Press Release, Marco Rubio: US Senator for Florida, Warner Express Concern over Growing Threat Posed by Deepfakes, Rubio (Oct. 2, 2019), https://www.rubio.senate.gov/public/index.cfm/2019/10/rubio-warner-express-concern-over-growing-threat-posed-by-deepfakes [https://perma.cc/QB98-C4AD].

it is clearly labelled as a fabricated video.[95] A Texas law makes the creation and distribution of a deepfake a misdemeanor if its intent is to influence an election.[96] While not specific to elections, Virginia has outlawed the use of deepfake technology to create nonconsensual pornography.[97] Other states have legislation pending. A Massachusetts bill would criminalize the use of deepfakes to further crimes or torts.[98] Another bill proposed in New York state would extend portrait protections for forty years after one's death.[99]

Beyond legislative solutions, some commentators have argued that existing law could be extended to hold the creators of deepfakes liable. If a deepfake involved copyrighted content, for example, the person who created the original content could sue the creator or otherwise make use of notice-and-takedown provisions to keep the video from spreading.[100] Additionally, several torts could provide the basis for suits against the creators of deepfakes. If a business created a deepfake purporting to show someone endorsing a product, the business could be sued in tort for violating one's "right of publicity," because they misappropriated an individual's likeness.[101] If a deepfake were particularly humiliating, its creator could be sued for intentional infliction of emotional distress.[102] Most pertinently, individuals could sue for defamation.[103] Public officials could sue if a creator acted with actual malice, and private individuals could sue if a creator was negligent.[104] Other torts rooted in privacy may also be relevant, though a suit's success would be doubtful.[105]

The reason many solutions to deepfakes focus on the liability of the creators is because Section 230 of the Communications Decency Act makes online platforms immune from liability for content posted on their sites.[106] Section 230 states that online platforms are not the speaker of material posted on their sites

[95] Sumagaysay, *supra* note 20.

[96] *Id.*

[97] Ferraro, *supra* note 91, at 2.

[98] *Id.* at 2–3.

[99] For the House Bill, see A08155B, N.Y. State Assemb., Reg. Sess., (N.Y. 2018), available at https://nyassembly.gov/leg/?default_fld=&leg_video=&bn=A08155&term=2017&Summary=Y&Text=Y [https://perma.cc/EX42-C9E7]; for the most recent version of the Senate Bill see S. 5959B, N.Y. State Assemb., Legis. Sess. (N.Y. 2019), https://www.nysenate.gov/legislation/bills/2019/s5959/amendment/b [https://perma.cc/97CJ-U5NV]; *see also* Ferraro, *supra* note 91, at 3 (discussing the New York bills).

[100] Chesney & Citron, *supra* note 5, at 1793.

[101] *Id.* at 1794; Russell Spivak,*"Deepfakes": The Newest Way to Commit One of the Oldest Crimes*, 3 GEO. L. TECH. REV. 339, 383–86 (2019).

[102] Chesney & Citron, supra note 5, at 1794; Douglas Harris, *Deepfakes: False Pornography Is Here and the Law Cannot Protect You*, 17 DUKE L. & TECH. REV. 99, 111, 113 (2019) (discussing both intentional and negligent infliction of emotional distress).

[103] Chesney & Citron, *supra* note 5, at 1793; Spivak, *supra* note 101, at 368–76.

[104] Chesney & Citron, *supra* note 5, at 1793–94.

[105] *See id.* at 1794–95 (surveying various privacy torts and outlining their limitations); Spivak, *supra* note 101, at 377–83 (discussing the same topics).

[106] *See* 47 U.S.C. § 230; Chesney & Citron, *supra* note 5, at 1795–98; Spivak, *supra* note 101, at 387–90; Vanessa S. Browne-Barbour, *Losing Their License to Libel: Revisiting § 230 Immunity*, 30 BERK. TECH. L.J. 1505, 1523–24, 1528 (2015).

by third parties and, thus, cannot be sued for ineffectively regulating content.[107] Congress passed Section 230 in response to a New York state court case, *Stratton Oakmont, Inc. v. Prodigy Services Co*.[108] The case is explicitly cited in the Congressional Record of Section 230's passage.[109] In *Stratton Oakmont*, the plaintiff sued Prodigy, which operated an online bulletin board for the radio show "Money Talk."[110] In October of 1994, an unknown user posted on the bulletin board claiming that Stratton Oakmont's President had committed criminal and fraudulent acts, in addition to saying that the firm was a "cult of brokers who either lie for a living or get fired."[111] Stratton Oakmont sued Prodigy for libel and sought partial summary judgment on the question of whether Prodigy was a "publisher" of the libelous statements.[112] The court concluded that Prodigy was a publisher and was liable to Stratton Oakmont for damages based upon any successful libel claim.[113] The court based its holding, in part, on Prodigy's decision to regulate the content posted on the bulletin board: because Prodigy exercised editorial control for its own benefit, it also had to accept the corresponding liability for what was posted.[114] Subsequently, Congress, not wanting new internet companies to be liable for content on their websites, passed Section 230.[115]

Section 230 has been interpreted to grant broad immunity to internet service providers and websites for content posted by third parties.[116] Due to the breadth of this immunity, several commentators have argued in favor of Section 230's amendment. Some argue that immunity should either be removed for bad actors or conditioned on the exercise of a reasonable standard of care.[117] Under this approach, companies protected by Section 230 would be stripped of their immunity if they "knowingly or intentionally," allowed third parties to post illegal or injurious content.[118] Alternatively, a platform would be protected by Section 230 only if it could prove that it had "take[n] reasonable steps to prevent or address unlawful uses of its services."[119] Immunity would be conditioned on a

[107] 47 U.S.C. § 230(c); Browne-Barbour, *supra* note 106 at 1528; Chesney & Citron, *supra* note 5, at 1797–98.

[108] Browne-Barbour, *supra* note 106, at 1518–19.

[109] *Id.* at 1519. Congress was already debating the amendment when the decision in *Stratton Oakmont, Inc.* was released. Stratton Oakmont v. Prodigy Servs. Co., 1995 N.Y. Misc. LEXIS 229, at *14 (N.Y. Sup. Ct. May 24, 1995).

[110] *Stratton Oakmont*, 1995 N.Y. Misc. LEXIS 229, at *3.

[111] *Id.* at *1–2.

[112] *Id.* at *2.

[113] *Id.* at *1.

[114] *Id.* at *13–14.

[115] *See* Browne-Barbour, *supra* note 106, at 1526.

[116] *Id.* at 1527–28.

[117] Danielle Keats Citron & Benjamin Wittes, *The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity*, 86 FORDHAM L. REV. 401, 418–19 (2017) (arguing either for an amendment that excepts the "worst actors" from immunity or a general conditioning of immunity on reasonable care); Danielle Citron & Quinta Jurecic, *Platform Justice: Content Moderation at an Inflection Point*, HOOVER WORKING GRP. ON NAT'L SEC., TECH. & L., 1, 8–10 (Sept. 15, 2018), https://www.lawfareblog.com/platform-justice-content-moderation-inflection-point [https://perma.cc/F8ZN-RCPF] (arguing similarly for a bad-actors exception and reasonable standard of care).

[118] Citron & Jurecic, *supra* note 117, at 8–9; *see also* Citron & Wittes, *supra* note 117, at 419.

[119] Citron & Wittes, *supra* note 117, at 419.

platform showing that its response to illicit content on their site was reasonable.[120] The latter of these proposals has been applied to deepfake videos on social media.[121] Specifically, the reasonable standard of care would be applied to companies' responses to deepfake videos posted on their websites, but it would also include a sunset provision with: a data-gathering requirement; damage caps; an anti-SLAPP provision; and an exhaustion-of-remedies provision requiring potential litigants to provide notice to the platform, after which the platform could review the material and decide whether to remove it.[122] Another solution, not specific to deepfakes, would amend Section 230 to resemble provisions in the Digital Millennium Copyright Act, making online platforms liable for defamatory material posted on their sites if the platforms did not remove the material after sufficient notice.[123] While not an explicit amendment, a narrower judicial construction of Section 230 may be sufficient to allow for defamation suits, even as the law is currently written.[124]

Some solutions have focused on the role of government regulatory agencies in regulating deepfakes. One proposal would allow the Federal Trade Commission ("FTC") to fine certain promulgators of fake news,[125] but only if they distributed fabricated news articles for financial gain.[126] For a broader enforcement scheme, Section 230 would have to be amended alongside regulatory changes.[127] In general, most proposals assume that the FTC would be the regulatory agency to enforce any legislation or regulation targeting deepfakes.[128] Other agencies, such as the Federal Communications Commission, have also been considered.[129] Due to Section 230, however, little attention has been given to the role of federal regulatory agencies in stopping the spread of deepfakes.[130] While many approaches have been considered for regulating the spread of deepfakes, none present a unified mechanism for solving the problem; the current patchwork landscape of deepfake regulation is not sufficient to prevent the videos' spread, and each solution has substantive limitations.[131]

[120] *Id.*

[121] Chesney & Citron, *supra* note 5, at 1799.

[122] *Id.* at 1800.

[123] Browne-Barbour, *supra* note 106, at 1554–55.

[124] *Id.* at 1551–53.

[125] *See* John Roberts, Note, *From Diet Pills to Truth Serum: How the FTC Could Be a Real Solution to Fake News*, 71 FED. COMMC'NS. L.J. 105, 106 (2018); *see also* John Allen Riggins, Comment, *Law Student Unleashes Bombshell Allegation You Won't Believe!: "Fake News" as Commercial Speech*, 52 WAKE FOREST L. REV. 1313, 1314 (2017).

[126] Roberts, *supra* note 125, at 119.

[127] *See* Chesney & Citron, *supra* note 5, at 1806.

[128] *Id.* at 1804–06; *see also* Elizabeth Caldera, Comment, "*Reject the Evidence of Your Eyes and Ears:" Deepfakes and the Law of Virtual Replicants*, 50 SETON HALL L. REV. 177, 194 (2019).

[129] *See, e.g.*, Chesney & Citron, *supra* note 5, at 1806–08; Caldera, *supra* note 128, at 195–96 (discussing possible regulation by the FCC and the potential for an entirely new agency).

[130] *See, e.g.*, Chesney & Citron *supra* note 5, at 1804 (concluding that regulatory agencies have a limited role in regulating deepfakes).

[131] *See infra* notes 168–73 and accompanying text.

III. Analysis: No Silver Bullet–Policy and Constitutional Limitations to Successful Deepfake Regulation

The current legal landscape places the onus on victims to sue the creators of deepfakes, even though victims are not in an adequate position to prevent the spread of the videos.[132] Alternatively, the law gives discretion to companies to determine how to prevent the harms of deepfakes, but it provides no incentive for companies to regulate themselves.[133] Even though some legislative action has been taken, it is not particularly aggressive and only applies in individual states. Existing legislation and categories of civil liability focus entirely on videos' creators, with no responsibility imposed on social media companies.[134] Emphasizing the role of deepfake creators has led to solutions that are inapplicable to foreign actors and inefficient at deterring the circulation of the videos.[135] Ignoring the role of online platforms ensures that social media companies will not have a uniform or effective response to any widely circulated deepfake.[136]

Proposed solutions, while better than the status quo, still largely fail to address the deepfake threat.[137] Many of them merely provide new civil causes of action to victims, replicating the problems inherent with current civil liability regimes.[138] This is true even of amendments to Section 230, which would only allow new kinds of civil suits.[139] Neither criminal penalties nor civil suits would be sufficient to deter the creators of deepfakes, nor would they deter online platforms from acquiescing in the videos' circulation.[140]

A. Existing Criminal and Civil Liability Fail to Deter Actors and Leave Responsibility to Corporations

The law as it stands uses a combination of criminal penalties and extension of civil liability to deter deepfake actors. Since the federal government has yet to act, these solutions have been implemented by states.[141] An existing Texas law, which makes it a misdemeanor to distribute deepfake videos prior to an election,[142] is the best example of the criminal approach, although the proposed federal criminalization of deepfakes is significant as well.[143] Conversely, California's law, which would allow victims of deepfake videos to sue the videos' creators for economic and noneconomic damages, is emblematic of the civil liability approach.[144]

[132] *See infra* note 177 and accompanying text.
[133] *See infra* notes 168–78 and accompanying text.
[134] *See supra* notes 96–101 and accompanying text.
[135] Chesney & Citron, *supra* note 5, at 1792, 1804.
[136] *See infra* notes 170–78 and accompanying text.
[137] *See infra* section III.B.
[138] *Id.*
[139] *Id.*
[140] *See* discussion *infra* Section III.A.
[141] *See supra* notes 89–101 and accompanying text.
[142] Sumagaysay, *supra* note 20; Ferraro *supra* note 91, at 14–15.
[143] *See supra* note 91 and accompanying text.
[144] Sumagaysay, *supra* note 20; Ferraro *supra* note 91 at 10–12.

Legislation that criminalizes certain uses of deepfakes faces several limitations inherent within criminal law. First, any criminal investigation would require prosecutorial decisions about the priority of an investigation and allocating resources.[145] In the past, both federal and state governments have deprioritized cybercrimes and failed to combat them: local law enforcement regularly fail to investigate cyberstalking, and, while federal prosecutors can be more aggressive, they currently lack the resources to investigate deepfakes.[146] Further, it is highly likely that many malicious actors who wanted to circulate deepfakes, especially in the context of elections, would be from foreign countries, making criminal penalties ineffective.[147] Even for domestic actors, if state and federal law enforcement lacked the resources to pursue and prosecute deepfake creators, the mere existence of a law criminalizing the behavior would not deter them.[148] Beyond enforcement, penal solutions are susceptible to traditional problems associated with attempting to curb a behavior by criminalizing it: knowledge of the specific law, subjective perception of cost-benefit analysis, and general irrational action.[149] Thus, there are a variety of *ex-ante* and *ex-post* problems inherent in relying on criminal law as the primary solution to deepfake videos.[150]

Solutions relying on civil liability present similar problems. Like criminal penalties, civil suits would be unlikely to deter foreign actors, as they could not be brought to American courts and forced to pay damages.[151] This inherent problem of liability is compounded by the unique threat of deepfakes. Online platforms are global in nature, and many of the serious disinformation campaigns are being waged by foreign actors.[152] The inability to hold these foreign actors accountable would leave some of the most serious threats of deepfakes unresolved.[153] Additionally, a successful civil suit against a deepfake creator would, naturally, have to identify the creator. Given the nature of technological evolution, this is no easy task.[154] Creators could mask their location using technologies like Tor, which can make it virtually impossible to locate an IP address.[155] It could also be difficult to locate the original poster, since many deepfake videos would be shared widely across different platforms.[156] In a system relying on civil liability, the victim of a deepfake would bear sole responsibility for identifying

[145] Chesney & Citron, *supra* note 5, at 1801.

[146] *Id.*

[147] *Id.* at 1804.

[148] *See* Paul H. Robinson & John M. Darley, *Does Criminal Law Deter? A Behavioral Science Investigation*, 24 OXFORD J. LEGAL STUDIES 173, 173 (2004) ("Having a criminal justice system that imposes liability and punishment for violation deters. Allocation of police resources or the use of enforcement methods that dramatically increase the capture rate can deter. But *criminal law*—the substantive rules governing the distribution of criminal liability and punishment—does not materially affect deterrence . . . .").

[149] *Id.* at 174.

[150] *See, e.g.*, Brown, *supra* note 7, at 38–39.

[151] Chesney & Citron, *supra* note 5, at 1792.

[152] *Id.*

[153] *See* Brown, *supra* note 7, at 41 (discussing the jurisdictional problems of using civil liability to hold foreign actors accountable).

[154] Chesney & Citron, *supra* note 5, at 1792; Brown, *supra* note 9, at 40–41.

[155] Chesney & Citron, *supra* note 5, at 1792.

[156] *See id.*

and locating a video's creator, a task that is difficult at best and impossible at worst.[157] Even if creators could be identified, and even if that creator were in the United States, the victim would bear the expense of initiating a suit and may even choose not to initiate one due to embarrassment.[158]

The problems of both civil and criminal liability are compounded as long as the federal government forgoes enacting new legislation or regulations. Without federal government action, the states are left to develop their own solutions, creating a patchwork approach that breeds uncertainty and lacks power. Some states may refuse to adopt any solutions to the problem of deepfakes, and those that do may adopt legislative and regulatory regimes with differing degrees of aggression.[159] This could negate the deterrent effect of particular policies and complicate any uniform response to a deepfake that might be successfully instituted nationwide. Further, some states may adopt broader definitions of deepfakes than others or apply their laws in a greater number of instances that states with narrower laws. Differences based on definition and scope could lead to increased litigation or possibly spurious litigation,[160] decreasing the already moderate deterrent effect of state-by-state legislation. Even if all states passed deepfake legislation, the subsequent regime would almost certainly lack comprehensiveness and simplicity, as each would define the problem differently and implement substantially different solutions.[161] Each of these problems could complicate enforcement and make deterrence less effective. While action by the states certainly should not be discouraged, such action alone will not be sufficient to prevent the widespread circulation of deepfake videos.

Other than states, the actors who have taken the most steps to prevent the spread of deepfakes are online platforms and social media companies themselves.[162] Protected by Section 230 immunity, however, their efforts have been lackluster.[163] Response to edited videos has not been uniform. When the Pelosi "shallowfake" was being circulated across multiple platforms in 2019, each site had different responses: YouTube removed the video, Facebook merely directed users to reports calling it a fake, and Twitter did nothing.[164] When a similar video

[157] Brown, *supra* note 7, at 40–41.

[158] Chesney & Citron, *supra* note 5, at 1792–93; Brown, *supra* note 7, at 41.

[159] *See* Chyssa V. Deliganis & Steven P. Callandrillo, *Syringes in the Sea: Why Federal Regulation of Medical Waste is Long Overdue*, 41 GA. L. REV. 169, 203–09 (2006) (discussing various state regulations after the expiration of the Medical Waste Tracking Act).

[160] *See* Amy Lee-Goodman, *A "Natural" Stand Off Between the Food and Drug Administration & the Courts: The Rise in Food-Labeling Litigation and the Need for Regulatory Reform*, 60 B.C. L. REV. 271, 274 (describing how the FDA's refusal to define the term "natural" led to an increase in litigation).

[161] *See* Hayes Hagan, *How to Protect Consumer Data? Leave It to the Consumer Protection Agency: FTC Rulemaking as a Path to Federal Cybersecurity Regulation*, 2019 COLUM. BUS. L. REV. 735, 739–40 (2019) ("[State privacy laws] simply cannot provide the simplicity and comprehensiveness of federal regulation.").

[162] Cade Metz, *Internet Companies Prepare to Fight the "Deepfake" Future*, N.Y. TIMES (Nov. 24, 2019), https://www.nytimes.com/2019/11/24/technology/tech-companies-deepfakes.html [http://perma.cc/VT9R-P7QD].

[163] *See id.* (discussing problems companies are having detecting and removing deepfake videos); Jesse Lempel, *Combatting Deepfakes Through the Right to Publicity*, LAWFARE (Mar. 30, 2018, 8:00AM), https://www.lawfareblog.com/combatting-deepfakes-through-right-publicity [https://perma.cc/3ZQZ-DXV3].

[164] Kelly, *supra* note 2.

was circulated following Trump's 2020 State of the Union Address, neither Facebook nor Twitter removed the video.[165] Without some outside policy to incentivize a uniform response, it is likely that the major social media sites would respond to an actual deepfake in the same way. As both Pelosi videos show, this irregular response will do nothing to prevent a video from penetrating mainstream discourse.[166] The lack of an effective response to deepfakes is apparent from companies' current policy toward the videos. Twitter currently bans deepfake pornography and released a draft of its proposed deepfake policy, in which "synthetic or manipulated media" would be marked and people would be warned before sharing the content.[167] It would not, however, automatically remove it.[168] This "notice" practice was confirmed when the company released its official policy in 2020.[169] In contrast, Facebook has said it will entirely ban certain deepfakes[170] and has launched a "challenge" to develop new ways to detect manipulated videos.[171]

Thus, the status quo is failing to deal with deepfake videos. States are not positioned to be the primary actors creating solutions to deepfakes, and existing criminal penalties and legislative extension of civil liability have flaws that could neutralize their deterrence effect. Online platforms and social media companies, shielded with immunity, have been slow to develop responses. These problems are largely duplicated in the proposed solutions.

B. Proposed Legislation and Amendments to Section 230 Only Replicate the Problems of the Status Quo

Proposed solutions, whether they amend Section 230 or not, replicate the problems of existing policies because they do not change the actor who would

[165] Victoria Bekiempis, *Facebook and Twitter Reject Pelosi's Request to Remove Edited Trump Video*, GUARDIAN (Feb. 9, 2020, 11:59 AM), https://www.theguardian.com/us-news/2020/feb/09/nancy-pelosi-trump-state-of-the-union-video-twitter-facebook [https://perma.cc/2ELZ-HSAV].

[166] For example, President Trump ultimately shared both the 2019 and 2020 videos. Kelly, *supra* note 2; Bekiempis, *supra* note 165; Drew Harwell, *Faked Pelosi Videos, Slowed to Make Her Appear Drunk, Spread Across Social Media*, WASH. POST (May 24, 2019, 3:41 PM), https://www.washingtonpost.com/technology/2019/05/23/faked-pelosi-videos-slowed-make-her-appear-drunk-spread-across-social-media/ [https://perma.cc/2HHC-QYGD].

[167] Sarah Perez, *Twitter Drafts a Deepfake Policy That Would Label and Warn, But Not Always Remove, Manipulated Media*, TECHCRUNCH (Nov. 11, 2019, 9:11 AM), https://techcrunch.com/2019/11/11/twitter-drafts-a-deepfake-policy-that-would-label-and-warn-but-not-remove-manipulated-media/ [https://perma.cc/KRF5-PLTK].

[168] *Id.*

[169] Sarah Perez, *Twitter's Manipulated Media Policy Will Remove Harmful Tweets & Voter Suppression, Label Others*, TECHCRUNCH (Feb. 4, 2020, 3:00 PM), https://techcrunch.com/2020/02/04/twitters-policy-on-deepfakes-and-manipulated-media-will-only-remove-harmful-tweets-including-voter-suppression/ [https://perma.cc/UWN2-PUS3].

[170] Aaron Holmes, *Facebook Just Banned Deepfakes, but the Policy Has Loopholes– and a Widely Circulated Deepfake of Mark Zuckerberg Is Allowed to Stay Up*, BUS. INSIDER (Jan. 7, 2020, 4:07 PM), https://www.businessinsider.com/facebook-just-banned-deepfakes-but-the-policy-has-loopholes-2020-1 [https://perma.cc/WFD9-4Z94].

[171] Mike Schroepfer, *Creating a Data Set and a Challenge for Deepfakes*, FACEBOOK (Sept. 5, 2019) https://ai.facebook.com/blog/deepfake-detection-challenge/ [https://perma.cc/ML4L-26TR].

initiate proceedings against online platforms and deepfake creators.[172] Regardless of where liability is distributed, the party responsible for pursuing damages is the victim, and, with deepfakes specifically, the victim is rarely in the best position to enforce legislation.[173]

Proposed amendments to Section 230 still replicate many of the problems with civil liability.[174] Conditioning immunity on reasonable behavior would have the ultimate effect of providing victims with a civil cause of action against online platforms.[175] Even if the issues previously discussed could be resolved, proposed amendments to Section 230 have several problems. First, any successful civil action would have to quantify the harm done to the plaintiff to determine damages.[176] While this may not pose a problem in instances of economic impacts, in instances of election interference or civil unrest the damages would be difficult to quantify. In the election-eve scenario, how could one quantify the influence on the election? How could that influence be translated into money damages? Additionally, while the candidate would presumably have a cause of action if the deepfake portrayed him or her, what would the remedy be for individuals who relied on the deepfake and changed their vote or decided not to vote at all? What is the scope of the class of plaintiffs? Any solution to these questions would be seemingly arbitrary, and it is uncertain how it would unfold during litigation.

A notice and takedown provision, if the sole mechanism for preventing the circulation of a video, would have problems of its own. Like the reasonable standard of care proposal, a notice-and-takedown provision would still ultimately rely on victims to pursue civil causes of action.[177] This presents the same problems of quantifying damages and determining the scope of the class of plaintiffs.[178] Further, a notice-and-take down provision would have a temporal disadvantage: during the period when notice was given, and the material was reviewed, the video could continue to be circulated.[179] Additionally, determining whether a video should have been removed from a platform is not an easy task. As the varying platform responses with the Pelosi "shallowfake" demonstrate, reasonable actors might disagree on whether a video should be removed.[180] In the midst of that disagreement, the video could continue to garner attention and cause individuals to believe it is true.

[172] Chesney & Citron, *supra* note 5, at 1792–93 (discussing the problems associated with expecting victims to initiate suits); Brown, *supra* note 7, at 40–41 (addressing the same matter).

[173] Chesney & Citron, *supra* note 5, at 1792–93.

[174] *See supra* notes 151–58 and accompanying text.

[175] *See* Citron & Wittes, *supra* note 117, at 419 ("A modest alternative to a sweeping elimination of the immunity for state law would be to eliminate the immunity for the worst actors. . . . [T]he amendment could state 'Nothing in Section 230 shall be construed to limit or expand the application of civil or criminal liability . . . .'").

[176] 1 DAMAGES IN TORT ACTIONS (MB) § 1.01 (2019).

[177] Browne-Barbour, *supra* note 106, at 1554–55 (discussing an amendment to the CDA similar to the notice and takedown provision of the Digital Millennium Copyright Act that would allow individuals to seek redress under state common law defamation actions).

[178] *See* 1 DAMAGES IN TORT ACTIONS *supra* note 176 (discussing quantifying damages).

[179] Browne-Barbour, *supra* note 106, at 1555.

[180] *See supra* notes 164–66 and accompanying text.

The proposed deepfakes Accountability Act does not resolve the problems already outlined; it is a difference of degree, not kind. It chooses to hold creators solely liable and does not amend Section 230 in any capacity.[181] Its mandatory watermarking mechanism is supported by a criminal penalty and would allow victims to seek damages and injunctive relief.[182] While a combination of criminal and civil liability may be better than relying on one of the two, it does not solve the problems inherent with those regimes.[183] In particular, even if the two are combined, victims and the government have no redress against malicious foreign actors who distribute deepfakes.

Not only are there a myriad of policy difficulties associated with regulating deepfake videos, the First Amendment poses its own challenges.[184] Deepfakes constitute a kind of expression and, if the legislation is written too haphazardly, the solution risks being stricken down regardless of efficacy.[185]

C. Constitutional Challenges for Deepfakes Regulation

The First Amendment will shape the contours of any successful deepfake regulation because of the protections it affords to most speech.[186] Deepfake technology, as a means of manipulating videos, would be subject to First Amendment protections when the video content itself is otherwise protected by the First Amendment.[187] For this reason, most broad, content-based legislation would likely be unconstitutional, as would an outright ban on the technology.[188] This does not mean any regulation is doomed; it only means that deepfake legislation must be "carefully tailored . . . [to] certain harmful Deep fakes."[189] By focusing on those deepfakes which fall outside the scope of constitutionally protected speech, a new regulation would survive a First Amendment challenge.[190]

Harmful deepfakes, by their very nature, are not "truthful" speech– they are a kind of lie depicting something that did not happen and are presented to convince the public of their veracity.[191] The fundamentally false and counterfeit nature of deepfakes, however, does not automatically position them outside the

[181] Ferraro, *supra* note 91, at 7–9

[182] *Id.* at 7–8.

[183] *See supra* Section III.A.

[184] *See* Green, *supra* note 25, at 1445 (discussing the application of the First Amendment when regulating false political speech).

[185] *Id.* at 1462–63, 1448–49.

[186] U.S. CONST. amend. I; *see also* Holly Kathleen Hall, *Deepfake Videos: When Seeing Isn't Believing*, 27 CATH. U. J.L. & TECH 51, 62 (2018).

[187] Chesney & Citron, *supra* note 5, at 1790–91; *see also* Spivak, *supra* note 101, at 357–58 (discussing how the illegitimacy of content-based restrictions impacts possible deepfake regulation).

[188] Chesney & Citron, *supra* note 5, at 1791; Spivak, *supra* note 101, at 399; *see also* Ashcroft v. ACLU, 542 U.S. 656, 660 (2004) ("[C]ontent-based restrictions on speech [are] presumed invalid . . . the Government bear[s] the burden of showing their constitutionality.").

[189] Citron & Chesney, *supra* note 5, at 1791; *see also* Hall, *supra* note 186, at 62 (noting that any deepfake regulation focused on content would be subject to strict scrutiny).

[190] *See* Green, *supra* note 25, at 1486.

[191] *Id.* (discussing the harm of false campaign speech).

First Amendment's protections.[192] As the Supreme Court stated in *United States v. Alvarez*, "[a]bsent from those few categories where the law allows content-based regulation of speech is any general exception to the First Amendment for false statements."[193] Thus, falsity alone cannot be the basis of a regulation.

The most obvious way for a deepfakes law to survive a constitutional challenge would be for the law to target deepfakes which fall within an exception to the prohibition on content-based restrictions.[194] The most relevant exceptions are for statements intended to incite imminent lawless action, defamation and libel, and fraud.[195] Those false statements that receive First Amendment protection tend to be lies which have some kind of instrumental value—such as a journalist deceiving an organization about their identity—or are protected in order to guard against "indirectly chill[ing]" truthful speech.[196] Harmful deepfakes—those which threaten election integrity, economic stability, or public safety—would not seem to fall within any category of false statements protected by the First Amendment: they have no larger instrumental value, as their only intent is to deceive, and they contribute nothing to the larger marketplace of ideas.[197]

The task, then, is to classify the most threatening deepfakes into categories of unprotected speech. Determining the reason for deepfake videos' unprotected status will subsequently guide what kinds of regulation are admissible. Rather than trying to determine a single category that encompasses every threatening deepfake, the multifaceted nature of the deepfake threat will require multiple categories to be considered.

Deepfakes that seek to sow economic disinformation would fall within the fraud exception to the First Amendment.[198] Commercial speech, as a category, is protected under the First Amendment, and it must be evaluated by its specific content.[199] Even when commercial speech is protected, some kinds of regulation are still permissible; for example, regulations regarding time, place, and manner are allowed.[200] Additionally, commercial speech which is fraudulent does not enjoy any First Amendment protection.[201] Since fraud has no social value and

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

[193] *Id.*

[194] Green, *supra* note 25, at 1483–86 (discussing how a narrowly tailored law that targeted non-protected speech would survive a constitutional challenge).

[195] *Alvarez*, 567 U.S. at 717.

[196] Alan K. Chen & Justin Marceau, *High Value Lies, Ugly Truths, and the First Amendment*, 68 VAND. L. REV. 1435, 1437–38 (2015).

[197] Hall, *supra* note 186, at 64 ("The counterspeech doctrine has become unconvincing in light of the speed and efficiency with which false news travels, and with the inability of citizens to discern the truth in today's media landscape."); Erwin Chemerinsky, *Deepfakes Videos Threaten Our Privacy and Politics. Here's How to Guard Against Them*, SACRAMENTO BEE (July 13, 2019, 8:45 AM), https://www.sacbee.com/opinion/california-forum/article232515577.html [https://perma.cc/YRS6-P3DH].

[198] *See* Va. State Bd. of Pharmacy v. Va. Citizens Consumer Council, 425 U.S. 748, 761–62 (1976) (discussing the fraud exception to the First Amendment).

[199] *Id.* at 761.

[200] *Id.* at 770–71.

[201] Illinois *ex rel.* Madigan v. Telemarketing Assocs., 538 U.S. 600, 612 (2003); *see also Va. State Bd.*, 425 U.S. at 771.

can have victims, the government has a strong interest in preventing its dissemination and has substantial discretion in determining the proper regulation.[202] At the same time, the regulation cannot be so broad as to sweep in protected speech, in hopes of also capturing that which is fraudulent.[203]

Economic deepfakes that pose the greatest threat would clearly fall within the fraud exception. Almost by definition, deepfake videos are false and misleading in a way the Court has ruled to be unprotected.[204] In the case of a deepfake distributed the night before an IPO, or one purporting to show an executive or director acting outlandishly, the only conceivable purpose of the video would be fraudulent.[205] It would be a false statement designed to mislead investors, thereby manipulating the stock price, or to convince consumers to buy from a competitor.[206] This speech is clearly not protected by the First Amendment and could be regulated by the federal government.

Several mechanisms are already in place that could be used to regulate economic deepfakes. A deepfake in which a public figure endorses a product or business would easily qualify as false advertising within the scope regulated by the FTC.[207] More generally, the FTC's grant of jurisdiction "to prevent persons, partnerships, or corporations . . . from using . . . unfair or deceptive acts or practices affecting commerce," would seem to apply to harmful economic deepfakes.[208] Deepfakes which aim to alter consumer behavior by disseminating false information could be considered an unfair or deceptive trade practice.[209] Videos produced as part of a deepfake-for-hire disinformation campaign could thus be regulated by the FTC.[210] Further, the Securities and Exchange Commission ("SEC") already regulates and issues fines for internet fraud where a party distributes false information about stocks.[211] Between the jurisdiction of the FTC and current SEC fraud enforcement, federal administrative agencies already have many of the tools needed to regulate deepfakes without violating the First Amendment.

[202] Chen & Marceau, *supra* note 196, at 1444, 1447.

[203] *Madigan*, 538 U.S. at 619–20.

[204] *See id.* at 617 (discussing why false and misleading statements in the context of charitable solicitations are not protected).

[205] *See supra* text accompanying notes 63–66.

[206] *See supra* text accompanying notes 63–66.

[207] *See* Riggins, *supra* note 125, at 1324– 25.

[208] 15 U.S.C. § 45(a)(2) (2018).

[209] A Brief Overview of the Federal Trade Commission's Investigative, Law Enforcement, and Rulemaking Authority, F.T.C. (Oct. 2019), https://www.ftc.gov/about-ftc/what-we-do/enforcement-authority [https://perma.cc/Q3Y6-BGFV]; *FTC Policy Statement on Deception*, F.T.C. (Oct. 14, 1983), https://www.ftc.gov/system/files/documents/public_statements/410531/831014deceptionstmt.pdf [https://perma.cc/AP84-83QH]; *see also* Roberts, *supra* note 125, at 119.

[210] *See supra* text accompanying notes 67–71 (discussing the use of deepfakes in disinformation-for-hire campaigns)

[211] For a list of the kinds of internet fraud that the SEC regulates, see *Internet Fraud*, U.S. SEC (Feb. 1, 2011), https://www.sec.gov/reportspubs/investor-publications/investorpubscyberfraudhtm.html [https://perma.cc/749S-4N3Z].

To regulate deepfakes designed to influence elections, a distinction must be drawn between deepfakes produced by foreign and domestic actors. The Federal Election Commission ("FEC") would seem to have jurisdiction to regulate harmful deepfakes that would interfere with an election.[212] With regard to interference by foreign actors, 52 U.S.C. § 30121(a)(1)(A) makes it "unlawful for a foreign national, directly or indirectly, to make a contribution or donation of money or other thing of value . . . in connection with a Federal, State, or local election."[213] A recent memo by FEC Chairwoman Ellen L. Weintraub interpreted this rule to apply to foreign citizens, government entities, and foreign corporations.[214] Further, the phrase "anything of value includes . . . the provision of any goods or services without charge or at a charge that is less than the usual and normal charge . . . ."[215] Thus, it would not be difficult to view deepfakes distributed by foreign actors as a kind of campaign contribution meant to assist one candidate while hurting another. At the very least, the FEC could expand its interpretation of Section 30121 to include the creation and distribution of deepfake videos by foreign agents.[216]

Regulating deepfakes produced by domestic actors, on the other hand, is more difficult because of the protections the First Amendment gives to political speech. The Supreme Court has noted that political speech "occupies the core of the protection afforded by the First Amendment,"[217] and applies strict scrutiny when it is hindered, upholding the law only if it is narrowly tailored to a compelling state interest.[218] While this test has been exacting on state laws targeting false campaign speech,[219].the issues presented by those cases are substantially different from the problem presented by crafting a narrowly tailored deepfake regulation. Specifically, the issue is whether the federal government can prevent the distribution of knowingly false information ahead of an election.[220] The FEC does not appear to list such an action as an election crime.[221] If the federal government, like the states, has a compelling interest in running elections free of

[212] *See* 52 U.S.C. § 30107(a) (2018).

[213] 52 U.S.C. § 30121(a)(1)(A) (2018).

[214] Memorandum from Ellen L. Weintraub on Draft Interpretive Rule Concerning Prohibited Activities Involving Foreign Nationals, 1–2 (Sept. 26, 2019), https://www.fec.gov/resources/cms-content/documents/mtgdoc_19-41-A.pdf [https://perma.cc/V9GQ-WA9A].

[215] *Id.* at 2 (quoting 11 CFR § 100.52(d)(1)).

[216] Perez v. Mortg. Bankers Ass'n, 575 U.S. 92, 101 (2015) (holding that agencies do not need to follow the Administrative Procedure Act's notice-and-comments procedures when issuing or amending an interpretive rule).

[217] McIntyre v. Ohio Elections Comm'n, 514 U.S. 334, 346 (1995).

[218] *Id.* at 347.

[219] *See, e.g.*, *id.* at 357 (holding an Ohio law prohibiting anonymous pamphleteering unconstitutional); 281 Care Comm. v. Arneson, 766 F.3d 774, 795 (8th Cir. 2014) (stating that Minnesota law prohibiting knowingly false speech about ballot initiatives did not meet the requirements of strict scrutiny).

[220] Green, *supra* note 25, at 1483–84.

[221] *See generally* FED. ELECTION COMM'N, FED. ELECTION CAMPAIGN LAWS (2019), https://www.fec.gov/resources/cms-content/documents/feca.pdf [https://perma.cc/K8XT-HT59].

fraudulent information, then a law prohibiting the knowing distribution of malicious deepfakes, with appropriate exceptions for satire and parody,[222] could conceivably survive a First Amendment challenge. At the same time, such a law would confront the previously discussed problem of prohibiting speech purely because it is false, a category of speech which sometimes has First Amendment protections[223] and could encounter the troubles associated with a nationalized libel law.[224] If the compelling interest exists, such a law may resemble the current California law, which prohibits the knowing or reckless creation and distribution of deepfakes.[225]

An alternative route, and one that is perhaps more promising, is for the FEC to enact a new regulation targeting the dissemination of libelous videos to influence an election.[226] Since these videos would almost certainly target political candidates, any regulation would be subject to the stringent standards outlined in *New York Times v. Sullivan*.[227] A statement concerning public officials would only lose First Amendment protection if it is made with "actual malice"—a statement made that one knows is false or with "reckless disregard" for its truthfulness.[228] While this standard can be difficult to prove, deepfakes targeting elections could be said to be, *ex-ante*, libelous.[229] By their very nature, deepfakes are not true. They are false statements designed to sway an election by injuring a candidate's reputation. This distinguishes deepfakes from criticism of public officials that may contain certain factual inaccuracies but nonetheless are protected by the First Amendment to facilitate robust public debate.[230] Applied to the election-eve scenario, for example, the video in question would obviously be libelous.[231]

Additionally, the FEC, and the federal government more generally, could regulate the dissemination of libel in the context of elections without explicitly nationalizing libel law.[232] The latter presents a difficulty because libel is regulated by states through their police powers, a power not allotted to the federal government in the Constitution.[233] Instead, the federal government could base a

[222] *See* Caldera, *supra* note 128, at 199–200; *see also* Hall, *supra* note 186, at 62.

[223] *See generally* United States v. Alvarez 567 U.S. 709, 718 (2012); Chen & Marceau, *supra* note 196, at 1437.

[224] Michael J. Mannheimer, *Equal Protection Principles and the Establishment Clause: Equal Participation in the Community as the Central Link*, 69 TEM. L. REV. 95, 141–42 (1996) (discussing the difficult of enacting a federal libel law).

[225] Chemerinsky, *supra* note 197. If the law were enacted by a state rather than the federal government, it would survive a First Amendment challenge. *Id.*

[226] This may require an amendment to the FEC's jurisdiction, as the FEC currently focuses primarily on financial issues and does not assert jurisdiction to regulate the truth of campaign speech. *See* 52 U.S.C. § 30106(b)(1) (defining the FEC's ability to make policy); *see also* Chesney & Citron, *supra* note 5, at 1807–08 (discussing limitations to FEC regulation of deepfakes due to jurisdiction).

[227] New York Times v. Sullivan, 376 U.S. 254, 279–80 (1964).

[228] *Id.*

[229] *See id.*; Chen & Marceau, *supra* note 196, at 1448.

[230] *See Sullivan*, 376 U.S. 254 at 280–82; Chen & Marceau, *supra* note 196, at 1448.

[231] *See supra* text accompanying notes 2–4.

[232] *See supra* note 226, *(*discussing limitations on the FEC's current jurisdiction).

[233] Mannheimer, *supra* note 224, at 141–42.

regulation of libelous material affecting elections on one of its enumerated powers.[234] One possible justification could come through a combination of the Election Clause[235] and the implied power of Congress to make regulations regarding federal elections under Article II of the Constitution.[236] Specifically, the Supreme Court has, in the past, spoken of the ability of the federal government to protect its own elections.[237] In *Ex parte Yarbrough*, the court noted that, because elections are central to a republic, the federal government must be able to protect elections from violence, corruption, and fraud.[238] Similarly, in *Burroughs v. United States*, the Court stated that Congress had the power to pass legislation to protect elections because of the need for self-protection.[239] This rationale would justify a deepfakes-specific regulation: since deepfakes are libelous, they threaten the integrity of elections and the federal government has a compelling interest in preventing their dissemination.[240]

Further, the federal government may be able to regulate libelous content pertaining to elections under the Commerce Clause.[241] Two possibilities exist under the Commerce Clause to regulate deepfakes which target an election: either through analyzing the economic effects of voting or imbedding an anti-deepfake regulation in a larger scheme that targets interstate commerce. The first option derives from post-New Deal Commerce Clause jurisprudence. In the line of cases from *United States v. Darby*[242] to *Perez v. United States,*[243] the Supreme Court held that Congress could regulate an activity if its overall effect has an economic impact, regardless of whether the activity itself was economic.[244] Per this reasoning, Congress could regulate elections, because voting had a substantial economic effect: voting determines who becomes president, the president shapes economic policy, and that policy affects the interstate movement of goods.[245] Deepfakes, by this reasoning, would have a downstream effect on interstate commerce, because the videos would partially determine the Presidency,

[234] *See id.* at 142.

[235] *See* U.S. CONST. art I, § 4, cl. 1. *See generally* Franita Tolson, *The Elections Clause and the Underenforcement of Federal Law*, 129 YALE L.J. FORUM 171, 171–72 (Nov. 18, 2019), https://www.yalelawjournal.org/forum/the-elections-clause-and-the-underenforcement-of-federal-law [https://perma.cc/6CJS-87GB] (discussing the ability of the federal government to regulate elections under the elections clause).

[236] U.S. CONST. art II., § 1, cl. 2; *see also* Dan T. Coenen & Edward J. Larson, *Congressional Power over Presidential Elections: Lessons from the Past and Reforms for the Future*, 43 WM. & MARY L. REV. 851, 887 (2002) (discussing the implied power to regulate elections in the context of Article II).

[237] *Ex parte* Yarborough, 110 U.S. 651, 657–58 (1884); Burroughs v. United States, 290 U.S. 534, 545 (1934).

[238] *Yarborough*, 110 U.S. at 657–58.

[239] *Burroughs*, 290 U.S. at 545. The court further noted that "Congress, undoubtedly, possesses that power [to safeguard elections], as it possesses every other power essential to preserve the departments and institutions of the general government from impairment or destruction, whether threatened by force or by corruption." *Id.*

[240] *See* Rebecca Green, *Counterfeit Campaign Speech*, 70 HASTINGS L. J. 1445, 1460 (2019) (arguing that the government has a compelling interest in regulating counterfeit campaigns speech because of the unique threat it poses to elections).

[241] *See* U.S. CONST. art. I., § 8, cl. 3.

[242] United States v. Darby, 312 U.S. 100, 118 (1941).

[243] Perez v. United States, 402 U.S. 146, 151–52 (1971).

[244] Coenen & Larson, *supra* note 236, at 879.

[245] *Id.*

in turn affecting the economy.[246] Of course, this argument would conflict with the recent turn away from the substantial-effect test that began with *Lopez v. United States*.[247]

Deepfakes could still be regulated under the modern Commerce Clause jurisprudence that arises after *Lopez*. In *Gonzales v. Raich*, the Court stated that Congress could regulate a noneconomic activity, if the "general regulatory statute [bore] a substantial relation to commerce . . . ."[248] If a regulation is part of a larger statute that concerns interstate commerce, then that regulation would fall within the scope of the Commerce Clause per the *Gonzales* court.[249] The most obvious candidate for such a regulation would be the Communication Decency Act, specifically Section 230.[250] The statute expressly defines its purpose as "preserv[ing] the vibrant and competitive free market that presently exists for the internet . . . [and] to encourage the development of technologies which maximize user control over what information is received . . . ."[251] The act clearly contemplated that the development of internet companies, and the dissemination of information on the internet, would have a substantial effect on interstate commerce.[252] Thus, inserting an election-specific regulation pertaining to deepfakes would be the sort of regulatory scheme endorsed by the *Gonzales* court.[253]

Finally, deepfakes that threaten public safety by inciting riots or causing mass panic would fall into the "imminent lawless action" exception applied in *Brandenburg v. Ohio*.[254] In *Brandenburg*, the Court determined that the First Amendment protected speech that merely argued for the moral imperative of violent action.[255] The court contrasted advocating the necessity of violence to accomplish a political goal,[256] with speech that advocated for violence *and* was likely to incite imminent lawless action.[257] The crucial distinction between protected and unprotected speech, then, is one of likelihood.[258] In *Brandenburg*, speech by a Ku Klux Klan member was protected because, while the member

[246] The President's ability to shape economic policy at the legislative level and regulatory policy within the executive branch links deepfakes with interstate commerce.

[247] *See* Lopez v. United States, 514 U.S. 549, 561 (1995); Coenen & Larson, *supra* note 236, at 880 (discussing the relation between *Lopez* and post-New Deal cases).

[248] Gonzalez v. Raich, 545 U.S. 1, 17 (2005) (quoting *Lopez*, 514 U.S. at 558).

[249] *See id.*

[250] *See* 47 U.S.C. § 230(a)(5) (2018).

[251] *Id.* § 230 (b)(2–3).

[252] *See id.*

[253] More broadly, 47 U.S.C. §§ 251–262 (2018) are concerned with the "Development of Competitive Markets," making it very plausible that all of Chapter 5, Section 200 of Title 47 is concerned with interstate commerce. Of course, this regulation is possible only if Section 230 immunity is amended.

[254] *Brandenburg v. Ohio*, 395 U.S. 444, 447 (1969).

[255] *Id.* at 448–49.

[256] *Id.* at 448.

[257] *Id.* at 447.

[258] *See id.* at 447, 448–49 ("[T]he constitutional guarantees of free speech and free press do not permit a State to forbid or proscribe advocacy of the use of force or of law violation except where such advocacy is . . . likely to incite or produce [imminent lawless action]. . . . Neither the indictment nor the trial judge's instructions . . . refined the statute's bald definition of the crime in terms of mere advocacy [nor] distinguished from incitement to imminent lawless action.").

talked about "reveng[e] in a way that alluded to violent action,"[259] violence was not likely enough for the speech to lose protection. Deepfake videos threatening public safety, however, are not merely a difference of degree from the speech in *Brandenburg*; they are a difference of kind.[260] For example, previous Russian disinformation campaigns have tried to convince the general population of chemical disasters or disease outbreaks.[261] If these campaigns were to be accompanied by a deepfake, the video would not be an abstract moral argument for the necessity of violence; indeed, its only purpose would be to sow panic and cause mass hysteria. It is conceivable that such videos could incite lawless action by breeding civil unrest.[262] For example, if a large contingent of the population believed a deadly disease was rapidly spreading in a metropolitan area, it is conceivable individuals would begin looting businesses for supplies. Likewise, if a deepfake were to be circulated during a tense protest, its only purpose, and likely effect, would be to turn the protest into a riot.[263]

In this way, a deepfake posted on social media to inflame tensions or manufacture public panic is the contemporary version of shouting fire in a crowded theater.[264] When such a deepfake is targeted in its distribution, the action meets the criterion put forth in *Brandenburg*: the threat it produces is both imminent and likely to produce lawless action.[265] A deepfake is clearly distinguishable from mere advocacy of the use of force or an abstract discussion of the moral necessity of violent action.[266] It is hardly even a statement of opinion. Instead, it is a tool used to inflame tensions that, when well placed, creates a high likelihood of chaotic, unlawful behavior. Clearly, this is the sort of speech intended to be excepted from First Amendment protection.

Of course, the mere possibility of regulation is insufficient if any federal action would only target private actors.[267] Ideally, any law would hold liable the medium through which deepfakes are distributed (*i.e.*, online platforms). It is impossible to hold platforms liable, however, if Section 230 continues to grant them immunity.[268] Amending Section 230 to hold platforms liable does not inherently present a constitutional challenge as platforms and Internet Service Providers ("ISPs") were originally considered publishers of material in *Stratton Oakmont, Inc. v. Prodigy*, which precipitated Congress's broad grant of immunity.[269] Thus, the current immunity is merely a Congressional preemption of a

[259] *See id.* at 446.

[260] Chesney & Citron, *supra* note 5, at 1803.

[261] *See supra* note 72 and accompanying text.

[262] Chesney & Citron, *supra* note 5, at 1803.

[263] *See supra* text accompanying notes 78–84.

[264] Schenck v. United States, 249 U.S. 47, 52 (1919) ("The most stringent protection of free speech would not protect a man in falsely shouting fire in a theatre and causing a panic.").

[265] *See Brandenburg v. Ohio*, 395 U.S. 444, 447 (1969).

[266] *See id.* at 447–48.

[267] *See supra* Section III.A.

[268] Brown, *supra* note 7, at 41–42.

[269] Stratton Oakmont v. Prodigy Servs. Co., 1995 N.Y. Misc. LEXIS 229, at *1 (N.Y. Sup. Ct. May 24, 1995).

judicial decision and could be congressionally overturned. The preemptive nature of Section 230 is evidenced by its phrasing, which states "[n]o provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider."[270] Congress could merely take away the immunity they previously granted.

Congress has previously amended Section 230 in a similar way. In 2018, Congress passed the Allow States and Victims to Fight Online Sex Trafficking Act ("FOSTA"), which revoked immunity for online platforms that knowingly supported or facilitated sex trafficking.[271] The passage of FOSTA both signals a political willingness to regulate online platforms[272] and additionally shows that immunity can be revoked without violating the First Amendment. Senator Rob Wyden of Oregon captured the compelling policy reason for amending Section 230 given the change in the internet ecosystem: Section 230 was originally enacted to incentivize companies to regulate content on their sites, not to encourage them to ignore illegality.[273] The extension of immunity to online platforms was a Congressional choice, and the decision to revoke it would not violate the First Amendment.

Simply rescinding Section 230 immunity, however, is not the end of the analysis. Even if immunity is taken away, it does not necessarily follow that social media platforms are the publishers of content posted to their sites. While Congress acted to preempt the *Stratton Oakmont* decision, the latter was a state court decision made in the context of a completely different internet.[274] Further, it is not clear to what extent a platform would be liable for what it is legally considered to have "published."[275] To answer these questions, the rationale of *Stratton Oakmont* must be applied to the contemporary internet.

The nexus of the *Stratton Oakmont* court's decision hinged on an analysis of "editorial control."[276] Indeed, editorial control was the "critical issue" in determining whether the defendant ISP was a publisher.[277] By emphasizing the agency of the ISP in editing its message boards, the court contrasted the defendant Prodigy with more basic computer bulletin boards: where the latter were mere repositories of information, Prodigy engaged in active curating of content posted to its site.[278] The decision to insert itself, to sculp the content according to its own guidelines, exposed it to the risk of liability—it could not gain the benefits of editorial control without the attendant costs.[279] This benefit was not merely abstract: Prodigy presented itself to the public as controlling the content of its

[270] 47 U.S.C. § 230(c)(1).

[271] Chesney & Citron, *supra* note 5, at 1798–99.

[272] *See* Citron & Jurecic, *supra* note 117, at 2.

[273] *See id.* at 2–3.

[274] *See generally* Stratton Oakmont v. Prodigy Servs. Co., 1995 N.Y. Misc. LEXIS 229, at *1 (N.Y. Sup. Ct. May 24, 1995).

[275] *See id.*; *see also* 47 U.S.C. § 230 (c)(1) (stating that internet providers shall not be considered "publishers" of information on their sites).

[276] *Stratton Oakmont*, 1995 N.Y. Misc. LEXIS 229, at *7.

[277] *Id.*

[278] *Id.* at *12–13.

[279] *Id.* at *13.

message boards, so that it could receive greater web traffic.[280] At the time of the case, Prodigy's computer network had two million subscribers, and its "Money Talk" bulletin board, where the comments giving rise to the suit were made, was the most read financial bulletin board in the United States.[281] Publication has benefits not associated with being a mere virtual library,[282] and the court did not allow Prodigy to reap the benefits of the former without incurring the penalties. Liability for harmful content, then, is a responsibility that must be borne along with the benefits of content-shaping.[283]

Not only does this rationale easily extend to social media companies, the underlying facts of *Stratton Oakmont* also bear a striking resemblance to current companies' content practices. In determining that Prodigy was a publisher, the court noted that it promulgated a set of content guidelines, screened all posting for offensive language with a software, employed "Board Leaders" to enforce the guidelines, and had a delete function where board leaders could immediately delete a post and send an explanation to its poster.[284] Contemporary social media companies employ many similar, and more advanced, practices, and they do so for the same reasons as Prodigy.[285] Facebook employs at least 15 thousand content moderators globally,[286] and has an extensive set of "Community Standards" that regulate online behavior such as statements that may incite violence, bullying and harassment, and violations of intellectual property law.[287] The stated purpose of these rules is to promote inclusivity, creative expression, and empower the voices of overlooked and marginalized communities, in addition to more general values like safety and privacy.[288] Facebook can remove individual posts and users.[289] Both Twitter[290] and YouTube[291] have similar policies and guidelines with similar purposes, as do nearly all major social media companies.[292] Social

[280] *Id.* at *10, *13–14 ("Presumably PRODIGY's decision to regulate the content of its bulletin boards was in part influenced by its desire to attract a mark it perceived to exist consisting of users seeking a 'family-oriented' computer service.").

[281] *Id.* at *3.

[282] *Id.* at *12.

[283] *See id.* at *13–14.

[284] *Id.* at *5–6.

[285] *See Id.* at *5–*11 (discussing Prodigy's content moderation policies and the reason for them).

[286] Casey Newton, *The Trauma Floor: The Secret Lives of Facebook Moderators in America*, VERGE (Feb. 25, 2019, 8:00 AM), https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona [https://perma.cc/MNK5-EKGL].

[287] *Community Standards*, FACEBOOK, https://www.facebook.com/communitystandards/introduction (last visited Jan. 15, 2021) [https://perma.cc/N48V-L9BQ].

[288] *Id.*

[289] *See id.*

[290] *See The Twitter Rules*, TWITTER, https://help.twitter.com/en/rules-and-policies/twitter-rules (last visited Jan. 15, 2021) [https://perma.cc/S8QV-FTUN].

[291] *See Rules and Policies: Community Guidelines*, YOUTUBE. https://www.youtube.com/about/policies/#community-guidelines (last visited Jan. 15, 2021) [https://perma.cc/GXV9-VC29].

[292] *See, e.g.*, *Community Guidelines*, INSTAGRAM, https://help.instagram.com/477434105621119 (last visited Jan. 15, 2021) [https://perma.cc/BC5F-MMBF]; *Community Guidelines*, TUMBLR, https://www.tumblr.com/policy/en/community (last visited Jan. 15, 2021) [https://perma.cc/D7YC-AZYC]; *Community Guidelines*, SNAP, INC., https://www.snap.com/en-US/community-guidelines (last visited Jan 15, 2021) [https://perma.cc/XB5Q-5QMY].

media companies are not merely passive repositories of information; they actively shape what is and is not allowable discourse.[293] Indeed, this content moderation is integral to their ability to attract users and remain central players in global discourse and communication.[294] It is hardly by accident that sites with robust content moderation polices are ubiquitous in every-day life, where completely unmoderated sites are confined to the periphery of the internet and overrun with abhorrent content.[295] Content moderation is a necessary component of how companies present themselves to the public, and it is indistinguishable from "publication" as understood by the *Stratton Oakmont* court.[296]

Holding publishers liable for content they disseminate is also a clear implication of the Supreme Court's ruling in *New York Times v. Sullivan.*[297] In *Sullivan*, the Court considered the Times' liability separately from the author of the advertisement published in the newspaper.[298] In considering the liability of the New York Times, the Court applied the actual malice standard separately from its application to the author.[299] This indicates that even if the author of an article—or, in this case, the creator of a deepfake—is liable for libel, a similar mental state must be found with respect to the publisher.[300] For the sake of argument, it is best to assume that the standard of "actual malice" would apply to the publisher of any deepfake video.[301] Actual malice is knowledge of the falsity of a statement, or reckless disregard for its truth,[302] and, in the case of deepfakes which become widely circulated, this standard will be met. If a deepfake were to become a part of wider public discourse, its existence would quickly become known to content moderators and officials at social media companies. This is true by definition: if a deepfake were broadly circulated, it would come to the attention of one of the many content moderators or executives employed by a social media company. If the video were not explicitly flagged, its popularity would lead many of the employees of the company to learn of it through their personal accounts. At the moment the company became aware of the video's existence, it would be faced with a choice: either act to remove all appearances

[293] Lee Rainie, Janna Anderson, and Jonathan Albright, *The Future of Free Speech, Trolls, Anonymity and Fake News Online*, PEW RSCH. CTR. (Mar. 29, 2017), https://www.pewresearch.org/internet/2017/03/29/the-future-of-free-speech-trolls-anonymity-and-fake-news-online/ [https://perma.cc/9U2K-SXJS] (discussing the effect of content moderation on internet discourse).

[294] *See id.*

[295] *See, e.g.*, Kevin Roose, *'Shut the Site Down,' Says the Creator of 8chan, a Megaphone for Gunmen*, N.Y. TIMES (Aug. 4, 2019), https://www.nytimes.com/2019/08/04/technology/8chan-shooting-manifesto.html [https://perma.cc/DQB6-2Y4H] (discussing the website 8chan).

[296] Stratton Oakmont v. Prodigy Servs. Co., 1995 N.Y. Misc. LEXIS 229, at *10–*11 (N.Y. Sup. Ct. May 24, 1995).

[297] New York Times v. Sullivan, 376 U.S. 254, 266 (1964).

[298] *See id.*, at 285–86 (considering the liability of the advertisement's author); *id.* at 286–88 (considering the liability of the New York Times for publishing the advertisement).

[299] *Id.* at 286.

[300] *See* Browne-Barbour, *supra* note 106, at 1553 (discussing the requisite mental state for ISPs to be liable for libel if Section 230 were not amended and the ISPs were only considered distributors).

[301] This is because "actual malice" is the most stringent standard, so if it can be met for the publication of any deepfake video, then liability will all but be assured in the scenarios within the scope of this paper.

[302] *N.Y. Times*, 376 U.S. at 280.

of the video on its platform or allow it to circulate. If the company were to choose the latter option, it would knowingly acquiesce in the dissemination of dangerous, unprotected speech, thereby making the company liable. In short, the very nature of the threat—broad public dissemination—ensures that the video's online saturation will reach a critical mass where its existence is widely known. At this point, a social media company's attention will be drawn to the video,[303] and a failure to remove it will impart liability. The point at which this critical mass is reached is ultimately a question of fact that would be determined by a regulatory agency if it chose to initiate an adjudicatory proceeding.[304]

IV. Recommendation: Using Federal Regulatory Powers to Hold Online Platforms Liable for Deepfake Videos

Given the necessity of federal action, and the insufficiency of private causes of action and criminal penalties, the government must use its regulatory powers to prevent the spread of deepfake videos.[305] Additionally, because holding individual creators liable is an inefficient means of deterrence, Section 230 should be amended to hold online platforms liable for content posted on their sites.[306] Specifically, Section 230 should be amended such that social media companies are treated as the publishers of deepfake videos.[307] The amendment should be specific to deepfakes and leave the broader immunity for ISPs in place.[308] It should also contain a watermarking provision, such that disclaimed deepfakes will be outside the scope of the amendment. In addition to the amendment, the necessary regulatory agencies should reinterpret their relevant regulations to include deepfakes,[309] or they should promulgate new rules as necessary.[310]

The overlapping spheres of administrative agency jurisdiction would allow multiple agencies to take part in regulating deepfakes. Indeed, this is the most preferable solution: rather than conceiving of deepfakes as a unified problem, they should be conceptualized as a multidimensional threat springing from a single technology.[311] Invoking multiple agencies would provide the best defense

[303] If the company were to fail to analyze the video, this would be reckless disregard for truth. Presumably, the video would be talked about by users and flagged by content moderators or machine learning algorithms. This would raise a reasonable suspicion that, if ignored, would be tantamount to a reckless disregard.

[304] Browne-Barbour, *supra* note 106, at 15532 ("[A] plaintiff must prove the distributor knew or reasonably should have known that the distributed material was defamatory."); Brown, *supra* note 7, at 50 ("[I]t is unclear when social platforms will be considered to have actual knowledge of deepfakes on their services.").

[305] *See* Caldera, *supra* note 128, at 193 ("[F]ederal administrative agencies would provide the fastest, most effective method of providing a form of regulation for deepfakes.").

[306] *See* discussion *supra* Section III.A.

[307] This would require amending 47 U.S.C. § 230(c)(1).

[308] For example, the amendment to Section 230 enacted by FOSTA left the larger grant of immunity intact. *See* 47. U.S.C. § 230(e)(5).

[309] If an agency were only to reinterpret an existing rule, it would not have to comply with the notice-and-comments requirements of the Administrative Procedure Act. Perez v. Mortg. Bankers Ass'n, 575 U.S. 92, 96–97 (2015); *see also* 5 U.S.C. § 553 (b)(A).

[310] If a new rule had to be promulgated, or an existing rule amended, the agency would have to comply with the standard procedural requirements for rulemaking. *See* 5 U.S.C. § 551(5); 5 U.S.C. § 553; *see also Perez*, 575 U.S. at 101.

[311] *See* Chesney & Citron, *supra* note 5, at 1771.

against the deepfake threat, because it would facilitate the sharing of information across departments and ensure no one agency is overburdened.[312] While there is no limit to the number of agencies that could regulate deepfakes, a few are immediately obvious. The FEC could issue fines for deepfakes which interfere in elections; the FTC and SEC could work in tandem to regulate economic deepfakes; and the Department of Homeland Security[313] could sanction deepfakes which pose a threat to public safety by inciting riots or panics.

Social media companies should specifically be held liable because they are in the best position to prevent the spread of deepfake videos.[314] Rather than criminalizing the creator or forcing the victim to seek redress on their own, extending liability to the medium in which the videos are circulated ensures they will be taken down by the platforms before they can be widely circulated.[315] The combination of regulatory enforcement and civil liability that would ensue from amending Section 230 would force companies to minimize a video's impact to avoid large fines or extensive litigation. Two policy goals are fulfilled by holding social media companies liable: deterring videos from being circulated and prompt removal if a video begins to be circulated.[316] To avoid liability, social media companies would invest in detection tools so that malicious deepfakes would be removed immediately upon being uploaded, and the companies would also monitor uploaded content so that a video would be removed if it began to attract significant attention.[317]

Regulatory liability would incentivize companies to take active measure to prevent deepfakes from becoming common on their sites. The possible fines the government could impose on social medial companies could be enormous, perhaps similar in size to the five billion dollars levied against Facebook for privacy violations.[318] These fines avoid the quantification problems associated with civil litigation, as the government has greater discretion to determine the size of the fine and negotiate a settlement with companies, although that discretion is not

[312] *See* Jason Marisam, *Interagency Administration*, 45 ARIZ. ST. L.J. 183, 190 (2013); *see also* Bijal Shah, *Congress's Agency Coordination*, 103 MINN. L. REV. 1961, 1968 (2019) (discussing the benefits of interagency coordination in the context of coordinating statutes).

[313] The Department of Homeland Security has the ability to issue fines in a variety of areas. *See, e.g.*, Civil Monetary Penalty Adjustments for Inflation, 84 Fed. Reg. 13,499 (Apr. 5, 2019) (to be codified at 6 C.F.R. pt. 27; 8 C.F.R. pts. 270, 274a, and 280; 19 C.F.R. pt 4; 33 C.F.R. pt 27; and 49 C.F.R. pt. 1503). Additionally, in 2018 the Homeland Security Act of 2002 was amended to include the Cybersecurity and Infrastructure Security Agency, which is charged with a variety of cybersecurity responsibility that could be relevant to deepfakes. *See* Cybersecurity and Infrastructure Security Agency Act of 2018, 115 Pub. L. No. 278, 132 Stat. 4168 (codified at 6 U.S.C. §§ 651–664). For a list of the CISA's responsibilities, see 6 U.S.C. § 652(c) and (e).

[314] *See* Danielle Keats Citron, *Cyber Civil Rights*, 89 B.U.L. REV. 61, 118 (2009) ("[B]road immunity for operators of abusive websites would eliminate incentives for better behavior by those in the best position to minimize harm.").

[315] This prediction is based upon the actions of social media companies following the European Union's decision to make them liable for certain kinds of extremist speech. Danielle Keats Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep*, 93 NOTRE DAME L. REV. 1035, 1037–38 (2018).

[316] *See id.* at 1048–49.

[317] *Id.*

[318] *See* Press Release, Fed. Trade Comm'n, FTC Imposes $5 Billion Dollar Penalty and Sweeping New Privacy Restrictions on Facebook (July 24, 2019), https://www.ftc.gov/news-events/press-releases/2019/07/ftc-imposes-5-billion-penalty-sweeping-new-privacy-restrictions [https://perma.cc/69LK-VUPN].

unlimited.[319] For example, agencies can consider the possible deterrence effect on a larger industry when deciding on the size of a fine,[320] while deterrence concerns in a civil proceeding are largely confined to discussions of punitive damages.[321] Conversely, administrative agencies can also channel retributive concerns if an offender acts with a particularly culpable mental state.[322] Thus, a company who acted particularly carelessly regarding a deepfake could be issued a higher fine.

Due to the sheer size of prospective fines, if instituted, regulatory liability would act as an economic deterrent. Penalizing actions that currently pose no financial risk to a company will cause that company, and other similarly situated parties, to change their behavior to minimize financial loss.[323] Liability alters a company's cost-benefit analysis: the benefits gained from engaging in a course of conduct must be weighed against the social harm of that conduct, and companies must internalize the cost of the harm.[324] For example, if a company is made liable for the effects of pollution, it is forced to weigh the benefit of not purchasing pollution equipment against the harm of pollution.[325] Similarly, in the case of deepfakes, companies will be forced to internalize the social harm of a widely distributed deepfake and invest the necessary resources to prevent the harm.

Further, whether a fine would ever be issued against social media companies is less important than influencing action by augmenting the shadow of the law, and the *possibility* of hefty penalties may be more important than any enforcement that would occur.[326] This has been the case in the European Union, where the threat of regulation has caused technology companies to adopt better safeguards against the proliferation of extremist speech.[327] In May 2016, Facebook, Microsoft, Twitter, and YouTube agreed with the European Union to remove hate speech within twenty-four hours of it being uploaded.[328] In response to the threat of penalties, the companies created an industry database of previously uploaded terrorist content, so that the content could be fingerprinted and removed if it were uploaded again.[329]

While a similar blacklist would be one response to the deepfake threat, wider regulatory liability would incentivize platforms to take a variety of cost-effective measures to prevent the spread of malicious deepfakes. For example,

[319] *See* Max Minzner, *Should Agencies Enforce?*, 99 MINN. L. REV. 2113, 2127 (2015); *see also* Walter Gellhorn, *Administrative Prescription and Imposition of Penalties*, 1970 WASH. U. L. REV. 265, 268 (1970) (stating that a legislature could not give unfettered discretion to an administrative agency to determine the size of a fine). Even though agencies are in theory constrained by statute in determining an appropriate fine, in practice they exercise a substantial amount of discretion in calculating a sanction. *See* Minzner, *supra*, at 2130.

[320] Minzner, *supra* note 319, at 2128–29.

[321] 4 DAMAGES IN TORT ACTIONS (MB) § 40.02[2] (2019).

[322] Minzner, *supra* note 319, at 2129.

[323] *See* A. Mitchell Polinsky & Steven Shavell, *Punitive Damages: An Economic Analysis*, 111 HARV. L. REV. 869, 877 (1998).

[324] *See id.* at 878.

[325] *Id.*

[326] *See* Citron & Jurecic, *supra* note 117, at 3–4.

[327] *Id.*

[328] Citron, *supra* note 315, at 1037–38.

[329] *Id.* at 1043–45.

companies could further invest in machine learning technology and artificial intelligence to identify deepfakes as these are uploaded.[330] These systems can be improved by analyzing the metadata associated with users to detect bot activity.[331] If a video was uploaded from a smartphone, companies could use technology to independently verify that the video was not manipulated before posting—if there is no verification, then the platform could automatically attach a disclaimer noting it is either unverified or likely fake.[332] Companies could also share detection systems with each other, so that the systems could advance and not be circumvented by new artificial intelligence programs used to create deepfakes.[333]

Beyond technological solutions aimed at deterrence, companies would also be incentivized to remove a video once it began circulating. Companies could institute notice-and-takedown provisions similar to those used in the enforcement of copyright law: companies could issue notices to users not to share videos and to remove the content if they had done so; if individuals refused to comply, their accounts could be deleted.[334] Beyond notifying the specific users, companies could increase the sophistication of content-flagging, such that users could flag videos specifically as deepfakes.[335] If a video was determined to be a deepfake, the site could simply remove the video or suspend the initial poster's account.[336] This could be coupled with additional human content moderators to review reported videos.[337]

Additionally, liability would ensure a uniform takedown response among platforms if a video began to circulate. The very nature of the threat means that a widely circulated video would become known to the platform on which it is

[330] This is how companies responded to European regulation. Citron & Jurecic, *supra* note 117, at 4. Generally, the use of AI and machine learning to identify deepfakes that are uploaded seems to be a primary tool companies are investing in to prepare for the threat. *See, e.g.*, Schroepfer, *supra* note 171 (describing the Facebook "Deepfake Detection Challenge" to use AI to identify videos edited using deepfake technology); Marie-Helen Maras & Alex Alexandrou, *Determining Authenticity of Video Evidence in the Age of Artificial Intelligence and in the Wake of Deepfake Videos*, 23 INT'L J. EVIDENCE & PROOF 255, 259–60 (2019) (surveying the state of AI and machine learning to detect deepfakes). *But see* Rini, *supra* note 5, at 7 (analyzing how the use of these technologies could create a technological arms-race with deepfake creators); John Villasenor, *Artificial Intelligence, Deepfakes, and the Uncertain Future of Truth*, BROOKINGS INST. (Feb. 14, 2019), https://www.brookings.edu/blog/techtank/2019/02/14/artificial-intelligence-deepfakes-and-the-uncertain-future-of-truth/ [https://perma.cc/9R6G-PEBR] ("Deepfake detection techniques will never be perfect. As a result, in the deepfakes arms race, even the best detection methods will often lag behind the most advanced creation methods."). For a general overview of the current state of deepfake-detection technology and its limitations, see Brown, *supra*, note 7, at 23–25.

[331] *See The National Security Challenges of Artificial Intelligence, Manipulated Media, and "Deep Fakes": Hearing Before the H. Permanent Select Comm. on Intel.*, 116th Cong. 7–8 (2019) (written statement of Jack Clark, Policy Director, OpenAI).

[332] *Id.* at 8.

[333] *Id.* at 8–9.

[334] Browne-Barbour, *supra* note 106, at 1547–51

[335] *See* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1638–39 (2018) (discussing the process of reactive content moderation).

[336] *See id.* at 1647–48 (discussing the process of removal and suspension and the options for appeal at Facebook, Twitter, and YouTube).

[337] *Id.* at 1639–41 (discussing the "tiers" of moderators used at Facebook).

shared.[338] With the knowledge that they would be liable for any impact of the video, they would take it down immediately upon learning of its existence. Any other response would expose them to a larger fine, and they would be unlikely to take the risk.[339] Regulatory liability would thus ensure a video was taken down before it had a significant impact. The deepfake would be caught in a double-bind: either the video would be shared among so few people that it wouldn't have a significant impact, or the video would be shared so often and so quickly that it would be detected and removed.

In the election-eve scenario previously described, deterrence and take-down provisions would work together to minimize the video's impact. Upon being uploaded, the video would be flagged by an algorithm designed to analyze recent content for deepfake video manipulation. The video would be reviewed by a human content moderator and, if it was not watermarked, it would be promptly removed. Depending on technological sophistication, the algorithm used to detect the video could be trained to detect watermarking on its own and to remove the video if it does not meet the requirement. The video would immediately be placed in a database shared with other social media sites, and, if it were uploaded on any other platform, it would immediately be flagged and removed. While the current response time from flagging by an algorithm to removal is several days,[340] increased liability will cause companies to invest in measures that would decrease the time from flagging to removal.

If the video managed to avoid detection by an algorithm, a concern of some commentators,[341] then efficient reactive measures would prevent its spread. If the video were not flagged by an algorithm, then, as the video began to circulate, a user could flag the video as a possible deepfake, triggering the removal process previously described. Since deepfakes would expose social media companies to a high degree of liability, then videos flagged as deepfakes would be prioritized for review along with other abhorrent content. If the video were not flagged, it would continue to be shared, but, at some point, the employees and executives of the social medial companies would encounter it on their personal feeds. Only one person would have to realize the video was a deepfake to begin the removal process. A concerted effort to remove the video would take place throughout election day. To mitigate any harm, the companies would likely release a statement saying the video is fake, along with media outlets and the government. The video would convince only those already susceptible to conspiratorial thinking. Confronted with multiple tiers of protection, the deepfake would be contained and largely removed from the sites before the end of voting. The story would not be about successful Russian interference, but attempted interference.

[338] *See supra* text accompanying notes 302–04.

[339] *See* Polinsky & Shavell, *supra* note 323, at 879–80 (discussing how, in a tort context, firms will take precautions if those precautions are less expensive than a prospective damage award).

[340] Brown, *supra* note 7, at 25.

[341] *See id.*

Some may argue that this proposal would verge into censorship and fail to distinguish between harmful and innocent content.[342] Regulations designed to prevent malicious deepfakes from large-scale distribution could become susceptible to "censorship creep" with effects beyond their original intention.[343] In analyzing the EU's agreement with social media companies, Danielle Citron notes three elements that make a policy ripe for censorship creep: definitional ambiguity, global enforcement of speech rules, and opacity surrounding private speech rights.[344] Each of these elements can be distinguished from the current proposal. The EU agreement suffered from definitional ambiguity because it targeted the broad category of "illegal hate speech."[345] The present proposal is much more narrowly tailored: it targets a specific kind of speech—videos created using deepfake or similar technology—and it only regulates those videos which are clearly malicious. The regulation would not ban the use of deepfakes for artistic or satirical purposes, and the watermarking provision would ensure innocent speech is protected.[346] Watermarking provides a clear bright line to distinguish illicit from allowed content.

Similarly, the regulation proposed in this Note is not as opaque as the EU agreement. It is true that companies will be removing content through a private, rather than court-supervised, process, raising some transparency concerns.[347] Companies can, however, combat this by releasing transparency reports detailing the number of videos removed, the number of requests for removal, and the process for analyzing videos.[348] Additionally, watermarking would help to reduce opacity, as it would provide an avenue for the use of deepfake technology without interference by social media companies. If companies and users can clearly delineate between permissible and impermissible uses of the technology, then opacity would be greatly reduced. Global enforcement and deletion, however, will be difficult to minimize. Global deletion will be an issue with the proposed regulations, as it was with the EU agreement, because the prohibition on malicious deepfakes will function similarly to the terms of service agreements that police illegal hate speech.[349] A concern with speed and uniformity may cause deletion of deepfake videos to become the default response.[350] This will not collapse into harmful censorship, however, because watermarking would allow users to use deepfake technology in a beneficial manner. By providing notice of new content policies, social media companies will provide space for users to change their behavior so that innocent speech will not be targeted.

[342] *See* Lauren Renaud, *Will You Believe It When You See It? How and Why the Press Should Prepare for Deepfakes*, 4 GEO. L. TECH. REV. 241, 255 (2019); *see also* Zeran v. American Online, Inc., 129 F.3d 327, 331 (4th Cir. 1997).

[343] Citron, *supra* note 315, at 1050–51.

[344] *Id.* at 1051.

[345] *Id.* at 1052.

[346] Caldera, *supra* note 128, at 199–200.

[347] *See* Citron, *supra* note 315, at 1057.

[348] *Id.*

[349] *Id.* at 1055.

[350] *Id.*

Other than censorship concerns, some critics may argue that rescinding Section 230 immunity would undermine the policy rationale that led to its passage: namely, that a lack of immunity would hinder the growth of technology companies and incentivize them to *decrease* the amount of editorial control they exercise over content.[351] While these concerns were justified in 1996, the evolution of technology companies over the last two decades makes this less of an issue in the current climate.[352] As the court noted in *Stratton Oakmont*, the fear that liability will cause companies to forego editorial control presumes that markets will not compensate companies for their "increased control and the resulting increased exposure."[353] The court noted that, in choosing to regulate the content posted on their sites, internet companies are able to market themselves to the public in a way that attracts users.[354]

This is even more true today than at the time of the *Stratton Oakmont* court's decision. Social media companies are dependent on active users to make money;[355] in Facebook's SEC filings, for example, it uses the metric of "average revenue per user" in analyzing its revenue.[356] A high number of users allows social media companies to make money through advertising and other activities that allow them to monetize the large cache of user data to which they have access.[357] In short, social media companies' profitability is directly tied to increasing the number of users on a site, and removing disturbing, violent, and obscene content is necessary to attract users and gain their trust.[358] The economic success of social media companies, then, is inextricably intertwined with their moderation of content. Social media and technology companies are no longer small businesses in need of government subsidization; they are multi-billion-dollar enterprises.[359] If they were to abdicate editorial control, they would risk losing large numbers of users and billions of dollars. When faced with regulation, the rational response would not be to forego content moderation; rather, companies would

[351] *See* Zeran v. American Online, Inc., 129 F.3d 327, 331 (4th Cir. 1997); Ryan M. Walters, *When Can You Shoot the Messenger? Understanding the Legal Protections for Entities Providing Information on Business Products and Services in the Digital Age*, 96 OR. L. REV. 185, 234 (2017) (discussing Congress's rationale for granting ISP's immunity).

[352] Omri Wallach, *How Big Tech Makes Their Billions*, VISUAL CAPITALIST (July 6, 2020), https://www.visualcapitalist.com/how-big-tech-makes-their-billions-2020/ [https://perma.cc/HV79-3EUY].

[353] Stratton Oakmont v. Prodigy Servs. Co., 1995 N.Y. Misc. LEXIS 229, at *13 (N.Y. Sup. Ct. May 24, 1995).

[354] *Id.* at *13–14.

[355] Brown, *supra* note 7, at 19.

[356] Facebook Inc., Annual Report (Form 10-K) 4 (Jan. 30, 2020), http://d18rn0p25nwr6d.cloudfront.net/CIK-0001326801/45290cc0-656d-4a88-a2f3-147c8de86506.pdf [https://perma.cc/4DY2-Q74T].

[357] Kalev Leetaru, *What Does it Mean for Social Media Platforms to "Sell" Our Data*, FORBES (Dec. 15, 2018, 3:56 PM), https://www.forbes.com/sites/kalevleetaru/2018/12/15/what-does-it-mean-for-social-media-platforms-to-sell-our-data/#53f097a72d6c [https://perma.cc/8FCQ-DN9N].

[358] Klonick, *supra* note 335, at 1627.

[359] For example, Instagram generated $20 billion in advertising revenue in 2019, and YouTube generated $15.1 billion in ad sales. Sarah Frier & Nico Grant, *Instagram Brings in More Than a Quarter of Facebook Sales*, BLOOMBERG (Feb. 4, 2020, 3:29 PM), https://www.bloomberg.com/news/articles/2020-02-04/instagram-generates-more-than-a-quarter-of-facebook-s-sales [https://perma.cc/5R32-ZQTC]; *see also* Citron & Wittes, *supra* note 100, at 421–22.

simply refine their content-moderation policies and practices to remain profitable. Arguments claiming social media companies would allow users to upload any and all content to their sites presuppose that a company would choose to implode in response to a narrow amendment of immunity.

Thus, basing deepfake regulation on an amendment to Section 230 as well as current jurisdictional grants to administrative agencies, would prevent the spread of deepfake videos. Regulatory liability would act as an economic incentive for platforms to prevent deepfake videos from being circulated and to remove videos before they would attract significant public attention. As was proven by a 2016 agreement with the EU, companies will respond to regulation by increasing their content moderation. A narrow, deepfake-specific amendment to Section 230 of the CDA would avoid problems associated with censorship, and companies would react to new regulation by increasing content moderation, rather than entirely foregoing it.

V. CONCLUSION

Deepfake videos pose an unprecedented threat to American democracy and national security. Already-existing foreign influence campaigns could use deepfakes to sow greater domestic discord,[360] undermine election integrity,[361] and cause disruptions to the economy.[362] Further, radical domestic actors could harness deepfakes to promote their own agendas, with many of the same effects.[363] Current solutions create a patchwork approach which fails to minimize the most harmful effects of deepfake technology.[364] Due to the failure of the status quo, Section 230 of the Communications Decency Act should be amended to treat online platforms as publishers of deepfakes posted on their sites by third parties, and federal agencies should subsequently reinterpret their regulations to encompass deepfake videos.

Social media companies can accurately be considered publishers because they exercise editorial control over the content posted on their sites, and in reap-

[360] *See* Chesney & Citron, *supra* note 5, at 1780, 1782 (describing deepfakes that could be introduced to inflame domestic tensions as well as the possible effect on Russian disinformation campaigns in Louisiana and Georgia); Atkinson, *supra* note 70 (detailing Russian disinformation campaigns during the Yellow Jacket Protests in France); *see also* Breland, *supra* note 79 (describing protests organized in the United States by Russian disinformation campaigns); Bloomberg, *supra* note 80 ("Russians accused of conspiring to help elect U.S. President Donald Trump allegedly staged rallies in key states to support him . . . .").

[361] *See* Chesney & Citron, *supra* note 5, at 1778–79 (analyzing how Russian election interference would be aggravated by the use of deepfakes); Citron, *supra* note 4, at 5 ("Imagine that the night before the 2020 election a deep fake showed a candidate in a tight race doing something shocking he never did. The deep fake, if spread widely, could alter the election's outcome."); *see also supra* note 58 and accompanying text (noting the possible use of deepfakes by China and Iran).

[362] *See supra* notes 9–67 and accompanying text.

[363] *See* Watts *supra* note 58, at 2 (noting the possible use of deepfakes by nonstate actors); Watts, *Advanced Persistent Manipulators*, *supra* note 60 (discussing the use of social media by extremist groups for disinformation purposes).

[364] *See supra* Section II.C.

ing the benefits of such control, they also take on the associated risk of liability.[365] Amending immunity and enacting subsequent regulation would not violate the First Amendment because the deepfake videos in question would fall into clearly recognized exceptions to the First Amendment.[366] Further, social media companies could be held liable because they would possess the necessary mental state. The very nature of the deepfake threat means that for a video to be distributed widely, the platform on which it was posted would be alerted to its existence. A failure to act would then be classified as knowing distribution of the deepfake or as reckless disregard.

New regulation would also incentivize technology companies to develop new techniques for removing deepfakes before they could be widely circulated. Companies could deploy machine learning technology,[367] user reporting,[368] and other mechanisms[369] to monitor their sites for deepfakes. Exposure to liability would force the companies to internalize the social harm posed by malicious deepfakes, and companies would develop strategies both to deter the spread of deepfakes as well as to remove them if posted. A regulation tailored specifically to deepfakes would avoid censorship concerns and would not cause companies to abdicate content moderation altogether.[370] Consequently, new federal regulations solve the most serious problems associated with a deepfake video entering into the larger public discourse.

[365] Stratton Oakmont v. Prodigy Servs. Co., 1995 N.Y. Misc. Lexis 229, at *13–*14 (N.Y. Sup. Ct. May 24, 1995).

[366] *See* United States v. Alvarez, 567 U.S. 709, 717–18 (2012) (discussing the categories).

[367] *See supra* note 330 and accompanying text.

[368] Klonick, *supra* note 335, at 1638–39.

[369] Clark, *supra* note 331, at 7–8.

[370] *See* Klonick, *supra* note 335, at 1627.