Regulatory Transparency Project
Unlocking Innovation & Opportunity

Deepfake Laws Risk Creating More Problems Than They Solve

Authored by:

Matthew Feeney

The Federalist Society and Regulatory Transparency Project take no position on particular legal or public policy matters. This paper was the work of multiple authors, and no assumption should be made that any or all of the views expressed are held by any individual author except where stated. The views expressed are those of the authors in their personal capacities and not in their official or professional capacities.

To cite this paper: M. Feeney, "Deepfake Laws Risk Creating More Problems Than They Solve," released by the Regulatory Transparency Project of the Federalist Society, March 1, 2021 (https://regproject.org/wp-content/uploads/Paper-Deepfake-Laws-Risk-Creating-More-Problems-Than-They-Solve.pdf).

Introduction

The term "Deepfake" describes audio and visual content created with adversarial deep learning techniques, specifically Generative Adversarial Networks (GANs). GANs are made up of a generator and discriminator designed to analyze data. Both train each other through a feedback loop.[1] The generator learns what kind of data best fools the discriminator, while the discriminator learns how to detect fake data. Such a technique allows for a variety of applications, one of the most notable of which is the development of Deepfakes.

Deepfake techniques allow users to make it appear as if someone said or did something they never said or did. Artists, documentarians, filmmakers, and many others have used Deepfakes to produce creative as well as potentially life-saving content. But like all technologies, Deepfakes can be used for harm, including assaults on people's dignity and media designed to interfere in elections and create political instability. Deepfake technology, like many other innovations before it, presents risks and opportunities.

Lawmakers and academics have proposed laws to mitigate the misuse. Many of these proposals take aim at specific uses such as election interference or revenge pornography.

Although Deepfake technology is relatively new, it is not clear that it creates unique problems. Accordingly, lawmakers and officials should proceed with caution when considering Deepfake technology. Absent careful consideration, legislation intended to prevent the malicious use of Deepfake technology could stifle its valuable uses.

I. We Have Been Here Before

Deepfake technology is new, but it does not present novel challenges. While it is undoubtedly true that Deepfake technology makes it much easier for content creators to make high quality forged media, lawmakers, security officials, and courts have had to address forged media before. Previous approaches to new technologies such as photography, video editing, and others can help inform our discussions about Deepfakes.

Since photography emerged in the 19th Century, society has developed public and private mechanisms to address its harmful uses, including the development of fake images. It did not take long after the development of photography for hoaxes and forgeries to emerge. In 1840, only two years after the first photograph featuring a human being, the French photographer Hippolyte Bayard developed the first intentional fake photograph.[2] The photograph was designed to make it appear as if Bayard had committed suicide. At this time, photography was primarily a skill set reserved to inventors.

A few decades later, the camera became a tool available to the amateur. In that year, George Eastman, a former bank clerk from Rochester, New York, released the Kodak #1 camera.[3] The

[1] (Overview of GAN Structure 2019, Zhang 2012, Overview of GAN Structure 2019) https://developers.google.com/machine-learning/gan/gan_structure
Ian Goodfellow et al., "Generative Adversarial Networks," June 2014, https://arxiv.org/pdf/1406.2661.pdf

[2] (Zhang 2012, Zhang 2012) https://petapixel.com/2012/11/15/the-first-hoax-photograph-ever-shot/

[3] (Fineman, Kodak and the Rise of Amateur Photography 2004) https://www.metmuseum.org/toah/hd/kodk/hd_kodk.htm

camera came loaded with a 100-exposure roll of film. Customers could send the entire machine roll back to the factory in Rochester, where new rolls were loaded and completed rolls were processed.[4] The increased use of cameras raised a wide range of legal, social, and political concerns.

In 1890, Samuel Warren and the future Supreme Court justice Louis Brandeis co-authored an article for the Harvard Law Review titled "The Right to Privacy."[5] In the article, Brandeis and Warren note that the common law has evolved to address intellectual and scientific developments, going on to argue that the rise of "instantaneous photographs" and newspaper gossip required protection of the "right to be left alone."[6] At a time when billions of people upload photos and videos of their homes, families, food, and social activities Brandeis' and Warrens' concerns about 19th Century cameras look quaint.

As cameras became more common, courts and lawmakers addressed privacy and dignity concerns associated with their use. In addition, technology developed to help journalists, intelligence agencies, and many others to identify fake photographs. Nonetheless, advances in photography editing services continued to prompt concern about widespread deception. In 1990, Newsweek published an article arguing that with the advent of "electronic photography" Chinese officials would be able to claim that authentic photographs of atrocities were faked.[7]

When considering the concerns associated with Deepfakes, we should keep past experiences in mind. Institutions have dealt with manipulated images in the past, and society is familiar with navigating changing attitudes about what is and is not private.

II. Abuses

Abusive deployments of Deepfakes range from the political interference to assaults on personal dignity. Such use of the technology has motivated most legislation designed to address Deepfake harms.

In politics, the ability to make it look as if a political ally or opponent said or did something they never did has advantages. Obvious applications come to mind, such as making it appear that your political opponent said something socially unacceptable, is suffering from a disability, or under the influence of an intoxicant.

Manipulated political images are not new. Authoritarian regimes have used altered photographs in attempts to erase dissidents and former allies from history.[8] Even in freer countries politicians, candidates, and their allies have used manipulated media to their benefit. For example, in 1950 Sen. Milard Tydings (D-MD) challenged claims Sen. Joseph McCarthy (R-WI) made about the number of communists working in the American government. During Tydings' reelection campaign, Sen. McCarthy's allies began circulating a photograph of Tydings meeting with U.S. Communist Party

[4] Ibid.

[5] (Brandeis 1890) https://www.jstor.org/stable/1321160?seq=1#metadata_info_tab_contents

[6] Ibid.

[7] Newsweek Staff, "When Photographs Lie," *Newsweek*, July 29, 1990. https://www.newsweek.com/when-photographs-lie-206894

[8] Oleg Yegrov, "How Stalin's Propaganda Machine Made People Vanish From Pictures," *Russia Beyond*, October 15, 2018. https://www.rbth.com/history/329317-stalin-propaganda-photos

leader Earl Browder.[9] The photograph was fake, and its circulation may have contributed to Tydings' defeat in the 1950 election. More recently, staff working for Sen. Marco Rubio (R-FL) during the Republican Party's 2016 primary rushed to inform reporters that a photo of Sen. Rubio shaking hands with President Obama, which appeared on the website therealrubiorecord.com, was fake.[10]

Other instances of manipulated media are less deceptive and cruder, seeking merely to alter a political opponent's appearance in an unflattering manner. In January 2019 Seattle's Fox affiliate aired footage of President Trump's Oval Office address about the planned wall on the southern border.[11] One of the affiliate's employees used Deepfake techniques to make the president's head appear larger than normal and his skin a bright orange hue.[12]

There are less obvious political applications of Deepfakes. For example, in February 2020, India's governing Bharatiya Janata Party (BJP) used Deepfake technology to alter campaign footage of Manoj Tiwari, a BJP member of India's lower legislative body. The video altered via Deepfake technology showed Tiwari speaking Haryanvi, a Hindi dialect. In the original video Tiwari spoke English. BJP altered the video so that the campaign message would be understood by as many potential voters as possible.[13] This is an example of political allies using Deepfakes to help the candidate who appeared in the original video.

Activists can use Deepfakes to make political figures appear to be allies of their cause. The environmental group Extinction Rebellion used Deepfake techniques to alter footage of a speech given by then-Belgian Prime Minister Sophie Wilmès. In the original speech, Wilmès addressed the COVID-19 pandemic. Extinction Rebellion activists altered the footage to make it appear as if Wilmès linked the cause of COVID-19 and the SARS and Ebola viruses to the "exploitation and destruction by humans of our natural environment."[14]

Even mere suspicion of the use of Deepfakes can have political implications. In October 2018, President of Gabon Ali Bongo Ondimba attended a summit in Saudi Arabia.[15] Later that month, Saudi media reported that Bongo had been hospitalized while at the summit. Bongo was not heard

[9] David Kaiser, "Ted Cruz is Not the First Politician to Cause Controversy With a Doctored Photo," *Time*, February 19, 2016. https://time.com/4231131/ted-cruz-tydings-browder-photo/

[10] Zeke J. Miller, "Rubio Campaign Fires Back at Cruz Over Photoshopped Image ," *Time*, February 18, 2016. https://time.com/4229092/marco-rubio-ted-cruz-photoshop/

[11] Kyle Swenson, "A Seattle TV station aired doctored footage of Trump's Oval Office speech. The employee has been fired," *The Washington Post*, January 11, 2019. https://www.washingtonpost.com/nation/2019/01/11/seattle-tv-station-aired-doctored-footage-trumps-oval-office-speech-employee-has-been-fired/

[12] Ibid.

[13] Charlotte Jee, "An Indian politician is using deepfake technology to win new voters," *MIT Technology Review*, https://www.technologyreview.com/2020/02/19/868173/an-indian-politician-is-using-deepfakes-to-try-and-win-voters/

[14] Gabriela Galindo, "XR Belgium posts deepfake of Belgian premier linking Covid-19 with climate crisis," *The Brussels Times*, November 9, 2020. https://www.brusselstimes.com/news/belgium-all-news/politics/106320/xr-belgium-posts-deepfake-of-belgian-premier-linking-covid-19-with-climate-crisis/

[15] Sarah Cahlan, "How misinformation helped spark an attempted coup in Gabon," *The Washington Post*, February 13, 2020. https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/

from for months.[16] Rumors spread that Bongo had suffered a stroke or had died. On the last day of 2018, Bongo reemerged and delivered an end of year address from Morocco.[17]

Something seemed amiss about the footage of Bongo's 2018 address. A neurologist suggested that Bongo had suffered a brain injury and had undergone cosmetic surgery. Others claimed that the footage had been changed with Deepfake technology. Two Deepfake detection tests found that the footage was likely authentic. Nonetheless, rumors related to the video only fostered political uncertainty in Gabon, and Gabonese soldiers attempted a coup in January 2019.[18]

The stories from India, Belgium, and Gabon provide a few examples of the impact the existence of Deepfakes can have on politics that should not go unappreciated: i) the use of Deepfakes to improve a political ally's image, ii) activists co-opting the image of a politician to help their cause, and iii) the risk of authentic videos being considered Deepfakes.

Abusive uses of Deepfakes are not reserved to politics. Perhaps the most notable use of Deepfakes is in the creation of "revenge pornography," a broad term that includes material designed to make it looks as if someone appeared in pornographic content when they did not. One of the earliest and most prominent public discussions about Deepfakes concerned the use of Deepfakes to make it appear as if A-list celebrities had performed in pornographic films.

Compared to most members of the public, A-list celebrities find it comparatively easy to demonstrate that such videos are not authentic. For those without fame and millions of dollars it is more difficult. Deepfake pornography can cause victims stress, anxiety, and incur significant social cost. After Indian journalist Rana Ayyub criticized the BJP's response to the rape of an eight-year-old Kashmiri girl, Deepfake pornography showing Ayyub's likeness began spreading. Ayyub's health suffered as a result of the subsequent stress related to the spread of the content.[19]

Deepfake pornography is not reserved to regular online pornography venues. One study found that 94 percent of Deepfake pornography is hosted on websites dedicated specifically to that kind of content.[20]

Deepfake technology is relatively new, but it fits comfortably into a pattern of media manipulation that is familiar.

In some cases, harms associated with Deepfakes will arise from mere accusations that content is a Deepfake rather than the use of Deepfake technology. In a world where Deepfake content is ubiquitous it is easier for someone to deny that genuine media is fake. This so-called "liar's dividend" could have as much a negative effect on politics as authentic Deepfakes. For example, in 2019 a sex tape featuring then Malaysian Minister of Economic Affairs Azmin Ali and a rival minister's male political aide emerged.[21] Same-sex acts are illegal in Malaysia, and Ali's allies

[16] Ibid.
[17] Ibid.
[18] Ibid.
[19] Rana Ayyub, "I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me: Rana Ayyub," Huffington Post, November 21, 2018. https://www.huffingtonpost.in/rana-ayyub/deepfake-porn_a_23595592/?guccounter=1
[20] The State of Deepfakes: Landscape, Threats, and Impact, Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen, September 2019. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf
[21] Ibid.

(including the prime minister) were quick to dismiss the tape as a Deepfake production, despite experts being unable to find evidence that it was.[22]

It is regrettable that same-sex acts are illegal in Malaysia. Yet even those who approve of criminalizing same sex conduct should be concerned about this controversy in Malaysia. Deepfake techniques can be used to make it appear as if someone has engaged in any matter of activity, and when authentic footage of illegal or inappropriate conduct emerges, we should be prepared for the subject of the footage and their allies to mount a "liar's dividend" defense. If the infamous footage of President Trump and Billy Bush had been released years later, rather than in 2016, it is likely that the president's allies would likely have claimed that the footage was a result of a Deepfake smear campaign.

III. Benefits

Despite the abusive potential of Deepfakes, they also have many valuable uses. Legislation designed to limit Deepfake abuses should be careful not to hamper such uses.

Perhaps the most newsworthy valuable use of Deepfakes is in satire and other First Amendment-protected content. Many content creators have uploaded Deepfake videos to platforms such as YouTube that highlight such uses. One video, created by the YouTube channel Ctrl Shift Face, shows an edited video of an appearance the comedian Bill Hader made on the NBC show Late Night with David Letterman. In the original video, Hader discussed the table read for the movie Tropic Thunder, performing impersonations of the actors Tom Cruise and Seth Rogan while doing so.[23] The creator of the Ctrl Shift Face channel took the original footage and used Deepfake technology to impose Cruise's and Rogan's faces onto Hader's.[24]

In October 2020, South Park creators Trey Parker, and Matt Stone released a comedy video titled "Sassy Justice," which featured widespread use of Deepfakes. The show featured Deepfake images of President Trump, Facebook CEO Mark Zuckerberg, and actress Julie Andrews.[25] Unsurprisingly, Parker and Stone did not exercise much restraint in using the technology to satirical effect.

The benefits of Deepfakes are not limited to satirical speech. Educators have also taken advantage of the technology by using it to create digital representations of dead historical figures. This allows for students and the general public to access content such as audio of President Kennedy delivering the speech he was scheduled to deliver in Dallas on the day of his assassination.[26] Samsung's AI and researchers at the Skolkovo Institute of Science and Technology have demonstrated Deepfake technology that brings still images such as the Mona Lisa and photos of Albert Einstein to life. Such applications have implications for educators as well as filmmakers.

Documentary filmmakers have already used Deepfakes in order to disguise those on film. In the 2020 HBO documentary "Welcome to Chechnya" filmmakers used Deepfakes to protect the

[22] Ibid.
[23] (Face 2019) https://www.youtube.com/watch?v=VWrhRBb-1Ig
[24] Ibid.
[25] Colby Hall, "Sassy Justice! South Park Creators Produce YouTube Show Featuring Deepfake of Trump … to Warn of Deepfakes," Mediaite, October 27, 2020. https://www.mediaite.com/trump/sassy-justice-south-park-creators-produce-youtube-show-featuring-deep-fake-of-trump-to-warn-of-deep-fakes/
[26] "JFK Unsilenced," Cere Proc. https://www.cereproc.com/en/jfkunsilenced

identities of gay Chechens whose sexual orientation can result in serious injury, torture, or death in their native Chechnya.[27] This allowed for the filmmakers to protect subjects' identities without having to resort to traditional anonymous interviews which feature the silhouettes of subjects. The result is a moving and compelling documentary that provides the viewer with an intimate insight into the lives of gay Chechens fleeing persecution. Without Deepfakes, such a documentary might be less compelling or gripping.

IV. Legislation

Lawmakers seeking to address the harms associated with Deepfake content should be aware of their benefits. Deepfake legislation that is too broad risks undermining valuable uses of the technology. In the past few years lawmakers at the state and federal level have proposed and, in some cases, passed Deepfake legislation. In most cases, this legislation would do more harm than good, and many of the proposals raise significant First Amendment concerns.

A. State Legislation

The Deepfake legislation enacted so far at the state level has fortunately been narrowly targeted to some of the specific harms outlined above. Under these laws, Parker and Stone's "Sassy Justice" and the HBO documentary about Chechen homosexuals would be legal. Yet these bills are not without risk of unintended consequences.

Lawmakers in California and Texas have passed laws seeking to address the use of Deepfakes being used for election interference.[28] Under the California law, it is illegal to create or distribute audio or video content showing political candidates altered to resemble real content within sixty days of an election. The Texas law is similar, though it applies to Deepfake content created without thirty days of elections. Lawmakers in Washington, Maine, and Maryland have also proposed Deepfake election legislation.[29]

The California and Texas laws deserve perhaps the most scrutiny given that they target political speech, arguably the most protected category of speech under the First Amendment. Laws that limit speech about politicians and candidates, including speech that ridicules or embarrasses them, should have to pass an exceedingly high bar, demonstrating that they are serving a compelling governmental interest. Both the California and Texas laws fail to do so.

California News Publishers Association Staff Attorney Whitney Prout described the California law as "unconstitutional."[30] The law does allow for exceptions, including for "satire or parody." Nonetheless, the law runs the risk of stifling valuable speech as broadcasters and many others reject

[27] Joshua Rothkopf, "Deepfake Technology Enters the Documentary World," *The New York Times,* July 1, 2020. https://www.nytimes.com/2020/07/01/movies/deepfakes-documentary-welcome-to-chechnya.html

[28] CA AB 730, TX SB 751

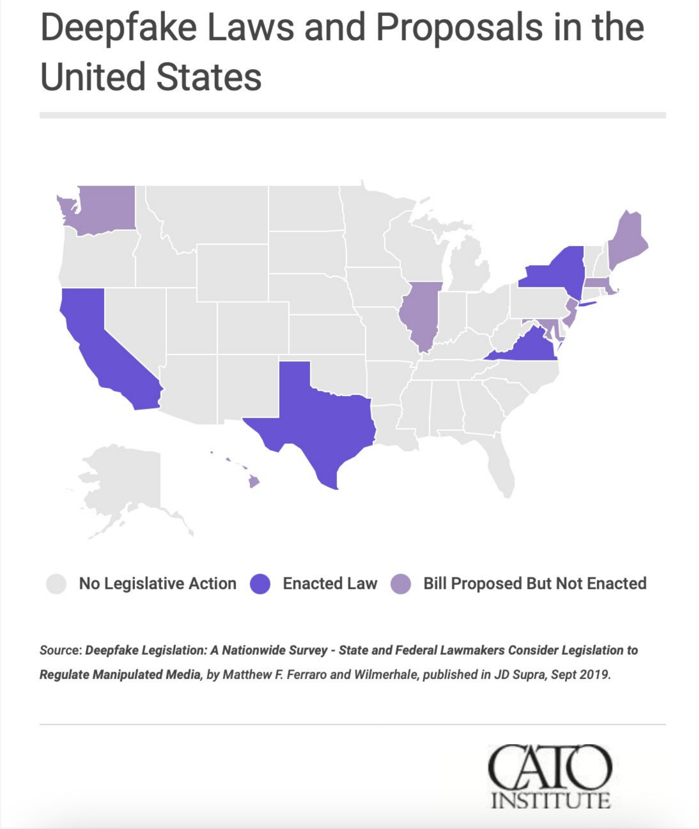
[29] HB 198, 2020 Regular Sess. (Md. 2020),
https://legiscan.com/MD/text/HB198/id/2100597/Maryland-2020-HB198-Introduced.pdf.
SB 6513, 2020 Regular Sess. (Wash. 2020),
http://lawfilesext.leg.wa.gov/biennium/2019-20/Pdf/Bills/Senate%20Bills/6513.pdf?q=20210209111130.
LD 1988, 129th Legislature, (Maine 2020)
http://www.mainelegislature.org/legis/bills/bills_129th/billtexts/SP069001.asp.

[30] Nick Cahill "Bill to Fight 'Deepfake' Videos Advances in California, Despite Free-Speech Fears," Courthouse News Service, July 2, 2019.
https://www.courthousenews.com/bill-to-fight-deepfake-videos-advances-in-california-despite-free-speech-fears/

political advertising and similar political content in an attempt to avoid liability. As Kevin Baker, the American Civil Liberties Union of California's legislative director, noted, "Despite the author's good intentions, this bill will not solve the problem of deceptive political videos; it will only result in voter confusion, malicious litigation, and repression of free speech."[31]

Figure 1, below, displays a map of current state laws and proposals regarding Deepfakes courtesy of Rachel Chiu of the Cato Institute.

Figure 1


Deepfake Laws and Proposals in the United States
● No Legislative Action
● Enacted Law
● Bill Proposed But Not Enacted
CATO INSTITUTE

*Source: **Deepfake Legislation: A Nationwide Survey - State and Federal Lawmakers Consider Legislation to Regulate Manipulated Media**, by Matthew F. Ferraro and Wilmerhale, published in JD Supra, Sept 2019.*

The Texas law is less nuanced than California's, neglecting to include exceptions for parody and satire.[32] Jared Schroeder, a journalism professor at Southern Methodist University, noted that the restrictions in Texas's law "cross free expression lines."[33] He went on to note that the Supreme Court has protected intentionally false speech.[34]

[31] Kathleen Ronayne, "California bans 'Deepfakes' video, audio close to elections," Associated Press, October 4, 2019. https://apnews.com/article/4db02da9c1594fd1a199ee0242c39cc2

[32] TX SB 751 https://legiscan.com/TX/text/SB751/id/2027638

[33] Jared Schroeder, "Texas deepfake law unlikely to survive scrutiny of the courts," *Texas Tribune,* September 13, 2019. https://blog.smu.edu/opinions/2019/09/25/texas-deepfake-law-unlikely-to-survive-scrutiny-of-the-courts/

[34] Ibid.

In July 2019, lawmakers in Virginia passed a law criminalizing the non-consensual sharing of Deepfake pornography. One analysis of Deepfake content found that the Deepfake pornography is the most prevalent kind of Deepfake material, far outpacing political Deepfake content.[35]

Virginia's law is narrow, prohibiting the sharing or creation of Deepfake pornographic content designed to "coerce, harass, or intimidate." The law also provides a liability protection for internet service providers. The narrow focus of the law and the liability protection are welcome, even if such liability protection is already provided under Section 230 of the Communications Decency Act. The law is relatively new, so it will take a few years to examine what unintended consequences arise. Nonetheless, its narrow focus on a specific category of content that is intended to do harm is an improvement on other state-level Deepfake proposals.

B. Federal Legislation

At the federal level, Sen. Benjamin Sasse (R-NE) is one of the lawmakers most concerned with Deepfakes. In the 115th Congress, Sen. Sasse introduced three Deepfake-specific bills.[36] These bills sought to address a range of harms.

Sen. Sasse's Malicious Deepfake Prohibition Act of 2018 would make it illegal to "create, with the intent to distribute, a Deepfake with the intent that the distribution of the Deepfake would facilitate criminal or tortious conduct under Federal, State, local, or Tribal law" and to knowingly distribute such Deepfake content.[37] The law also notes that interactive computer services could not be held liable for restricting access to Deepfakes, even though such content moderation is already protected under Section 230 of the Communications Decency Act.[38]

Two other Sen. Sasse Deepfake bills seek to institute studies of "cyberexploitation" that targets employees of certain federal agencies and their families.[39] Other legislation has established the study of Deepfakes. The IOGAN Act, signed into law by President Trump in December 2020, directed the Director of the National Science Foundation to "support research on the outputs that may be generated by generative adversarial networks, otherwise known as deepfakes."[40]

[35] The State of Deepfakes: Landscape, Threats, and Impact, Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen, September 2019.
https://storage.googleapis.com/deeptrace-public/Deeptrace-the-State-of-Deepfakes-2019.pdf

[36] S.3805 - Malicious Deepfake Prohibition Act of 2018, 115th Congress (2017-2018)
https://www.congress.gov/bill/115th-congress/senate-bill/3805/text
S.3788 - A bill to require studies on cyberexploitation of employees of certain Federal departments and their families, and for other purposes, 115th Congress (2017-2018).
https://www.congress.gov/bill/115th-congress/senate-bill/3788/text
S.3786 - A bill to require the Secretary of Defense to conduct a study on cyberexploitation of members of the Armed Forces and their families, and for other purposes. 115th Congress (2017-2018).
https://www.congress.gov/bill/115th-congress/senate-bill/3786/text

[37] S.3805 - Malicious Deepfake Prohibition Act of 2018, 115th Congress (2017-2018).
https://www.congress.gov/bill/115th-congress/senate-bill/3805/text

[38] Ibid.

[39] S.3788 https://www.congress.gov/bill/115th-congress/senate-bill/3788/text
S.3786
https://www.congress.gov/bill/115th-congress/senate-bill/3786/text

[40] U.S. Congress, Senate, Identifying Outputs of Generative Adversarial Networks Act (IOGAN ACT) Act of 2020, S 2904, 116th Cong., introduced in Senate November 20, 2019,
https://www.congress.gov/116/plaws/publ258/PLAW-116publ258.pdf

In 2019, Rep. Yvette Clarke, (D-NY) introduced a bill that would require those making Deepfake content to label the content so that viewers know the image has been altered.[41] Rep. Clarke's bill is an example of the broad legislation that lawmakers should avoid when seeking to address the spread of malicious Deepfakes.

The bill would impose an unnecessary burden on those creating First Amendment-protected media while doing little to deter motivated bad actors. Under Rep. Clarke's bill, those seeking to make satirical Deepfake content would have to include labels for such content. Numerous online video creators would be burdened with a requirement that applies to legal material. Perhaps more concerning is the fact that this burden's costs would be unlikely to be compensated with sufficient benefits. Watermarks and other labels attached to videos are relatively easy to remove. Those motivated to interfere with elections via Deepfake content would not find it difficult to circumvent the labeling requirement. Indeed, removing watermarks and identifying metadata can be automated. In addition, the presence of watermarks on media incurs a cost to content creators and their audience. As noted above, Deepfakes can be used in order to create valuable creative products. Part of the appeal of such products is the degree of realism and escapism Deepfake technology can provide. Watermark requirements or other labeling mandates would inevitably degrade the state of art, a significant cost that would come with little benefit.

Rep. Clarke's bill is an example of the kind of Deepfake bill lawmakers should seek to avoid. It is too broad, would impose a burden on those creating First Amendment-protected content, and would do little to mitigate the spread of damaging Deepfake content.

Some have argued in favor of amending existing law as a means to tackle Deepfakes. Professor Danielle Citron of Boston University's School of Law and Professor Robert Chesney of the University of Texas at Austin have proposed amending Section 230 of the Communications Decency Act as a means to mitigate the spread of harmful Deepfake content.[42] Citron and Chesney write:

> "Section 230 should be amended to allow a limited degree of platform liability relating to deep fakes. Building on prior work in which one of us (Citron) proposed a similar change in an article co-authored with Benjamin Wittes, we propose that Section 230(c)(1) protections to platforms be conditional rather than automatic. To qualify, an entity must demonstrate that it has taken "reasonable steps" to ensure that its platform is not being used for illegal ends. Platforms that meet this relatively undemanding requirement will continue to enjoy the protections of Section 230, but others will not and hence may be treated as a publisher of user-generated content that they host."[43]

Section 230 of the Communications Decency Act provides interactive computer services (such as social media sites) with two important protections, sometimes referred to as the "sword" and the

[41] H.R.3230 - Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019 116th Congress (2019-2020). https://www.congress.gov/bill/116th-congress/house-bill/3230/text

[42] Danielle K. Citron & Robert Chesney, Deepfakes: A Looming Challenge for Privacy, Democracy, and National Security, 107 California Law Review 1753 (2019). Available at: https://scholarship.law.bu.edu/faculty_scholarship/640

[43] Ibid.

"shield" of the law.[44] Section 230's shield provision, described as the "26 words that created the Internet," state that interactive computer services are not (with very limited exceptions) the publishers of content posted to such services by another user:

> "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider."[45]

If a Facebook user posts illegal content, the victim can pursue legal action against the user, but not (in most cases) Facebook. Section 230's sword states that interactive computer services are free to remove content they consider to be objectionable. This allows owners of websites big and small to take steps to remove content that might be legal but unconducive to the image a business might seek to present. Beheading videos and pornography might be legal, but it is not hard to see why a social media site might seek to distance itself from such content.

Citron and Chesney have proposed changing Section 230 to make its protections contingent on interactive computer services taking "reasonable steps" to prevent illegal activity. Their proposed amendment to Section 230's shield is below, with the suggested changes underlined:

> "No provider or user of an interactive computer service that takes reasonable steps to prevent or address unlawful uses of its services shall be treated as the publisher or speaker of any information provided by another information content provider in any action arising out of the publication of content provided by that information content provider."

As I have noted before, the inclusion of "reasonable steps" should give lawmakers tempted by such legislative changes pause.[46] In their paper, Citron and Chesney note that the amendment could come with a sunset provision and cite the renewal process of surveillance authorities under Section 702 of the US Foreign Intelligence Surveillance Act.[47] Yet the history of Section 702 reauthorization suggests that making Section 230 protections contingent on "reasonable steps" would prompt unhelpful rhetoric and commentary from lawmakers in Congress and market incumbents.[48]

In addition, Congress readdressing Section 230 protections every few years allows for anti-competitive behavior. Market incumbents will have an incentive to persuade lawmakers and regulators that their approach to Deepfakes is one that ought to be considered consistent with "reasonable steps." In such an environment, those seeking to compete with "Big Tech" could be put at a disadvantage given that they may not have access to Deepfake detection tools.

[44] 47 U.S.C §230

[45] Ibid.

[46] Matthew Feeney, "Defending the Indispensable: Allegations of Anti Conservative Bias, Deepfakes, and Extremist Content Don't Justify Section 230 Reform," CSAS Working Paper 20-11. https://administrativestate.gmu.edu/wp-content/uploads/sites/29/2020/03/Feeney-Defending-the-Indispensable.pdf

[47] Citron & Chesney

[48] Jake Laperruque, "Facts on FISA: Correcting the Record on the Section 702 House Floor Debate," Just Security, January 17, 2018. https://www.justsecurity.org/51110/facts-fisa-correcting-record-section-702-house-floor-debate/
Benjamin Wittes and Susan Hennessey, "Congress Wants to Tie the Intelligence Community's Hands for No Reason," Foreign Policy, October 13, 2017. https://foreignpolicy.com/2017/10/13/congress-wants-to-tie-theintelligence-communitys-hands-for-no-reason/

If Section 230 protections were made contingent on firms taking "reasonable steps" to address illegal content – including some Deepfake material – we should expect firms to engage in more aggressive content moderation. Some might argue that interactive computer services embracing false positives would be a net benefit to society, but we should be aware of the costs of such an approach, which would inevitably include valuable and First Amendment-protected content being stifled by firms seeking to avoid liability.

Legislative proposals to tackle Deepfakes suffer from a number of problems lawmakers should consider. Perhaps the most notable are the potential First Amendment concerns. Lawmakers should be hesitant to support legislation that could result in less speech, especially political speech. Even the Deepfake election law in California, which includes exceptions for satirical content and is restricted to content within sixty days of an election, raises significant First Amendment concerns.[49]

Broader legislation which seeks to mandate the labeling of Deepfake content would incur a significant creative cost while doing little to mitigate the most serious harms associated with Deepfakes. Legislation that seeks to amend Section 230 of the Communications Decency Act is potentially anti-competitive, as the largest incumbent firms would be best positioned for compliance and able to influence what lawmakers and regulators consider "reasonable steps" or best practices when it comes to Deepfakes.

Laws that target specific content that is intended to "coerce, harass, or intimidate," such as the Virginian Deepfake nonconsensual pornography law are unlikely to pose many of the issues outlined above, especially given that the law specifically exempts internet service providers from liability.

V. Private Solutions

When considering the risks associated with Deepfakes we should not limit ourselves to legislation and other government action. Private actors have developed methods for detecting Deepfakes.

This is unsurprising given that many online services have an incentive to ensure that users are not harassed or misinformed. We should expect such moderation efforts to yield some false positives and false negatives, especially given the scale of user generated content created by users of the most popular social media sites such as YouTube and Facebook. Although Deepfake detection methods will not be perfect, demand for them will increase as Deepfake content becomes ubiquitous.

Social media is not the only industry that has an incentive to halt the spread of misinformation. Journalism shares that incentive. If journalistic outlets consistently fail to identify Deepfakes, their reputations will suffer. In 2019, The Wall Street Journal formed a committee tasked with helping its newsroom identify fake content.[50] Reuters has also taken steps to address the rise of Deepfakes, as has The Washington Post.[51]

Journalistic and fact-checking outlets such as Animal Politico, Code for Africa, Rappler, and Agence France-Presse have used Assembler, a media manipulation detection tool built by Jigsaw, a Google

[49] First Amendment Watch, "California Becomes the Second State to Restrict Political "Deepfakes" ," October 9, 2019." https://firstamendmentwatch.org/california-becomes-the-second-state-to-restrict-political-deepfakes/
[50] Lucinda Southern, "'A perfect storm': The Wall Street Journal has 21 people detecting 'deepfakes'," Dig Day, July 1, 2019. https://digiday.com/media/the-wall-street-journal-has-21-people-detecting-deepfakes/
[51] "Seeing Isn't Believing: The Fact Checker's guide to manipulated video" *The Washington Post*. https://www.washingtonpost.com/graphics/2019/politics/fact-checker/manipulated-video-guide/

incubator.[52] Journalists and academics have collaborated in efforts to address the spread of Deepfakes, with Duke University's Reporters' Lab, the News Integrity Initiative at CUNY's Newmark School of Journalism, and Harvard's Nieman Lab being among the academic institutions seeking to help journalists tackle Deepfake material.

Attempts to identify Deepfake material will help household name firms in social media and journalism tackle the spread of misinformation and harassing content. Yet these detection tools can do nothing for those who do not wish to use them. Sadly there are websites dedicated to spreading Deepfake pornography and other manipulated content designed to harass and intimidate. Those creating such content might not be interested in Deepfake detection tools, but such tools will nonetheless prove valuable to those who wish to disassociate with Deepfakes and to safeguard their reputations. Those who are intent on using Deepfake techniques to harm others will not use these tools and will continue to motivate calls for legislation.

Conclusion

Harms associated with Deepfakes will continue to motivate lawmakers to propose legislation. When considering such legislation, lawmakers should consider that although Deepfakes technology is relatively new, it is not raising a host of unique challenges. Attempts to address Deepfake political interference and manipulated pornography should consider not only potential unintended consequences, such as the entrenchment of market incumbents, but also the effect such proposals might have on speech protected by the First Amendment. Legislation that seeks to target Deepfakes should be narrowly focused to a small category of content that seeks to inflict specific harms.

[52] Jigsaw, "Disinformation is More than Fake News," February 2020. https://medium.com/jigsaw/disinformation-is-more-than-fake-news-7fdd24ee6bf7