NOTES

DEEPFAKES, REAL CONSEQUENCES: CRAFTING LEGISLATION TO COMBAT THREATS POSED BY DEEPFAKES

Jack Langa[*]

ABSTRACT

Given the rapid development of deepfake technology, there is a growing consensus that its potential to cause deleterious consequences justifies some form of regulation. Proposed regulations are diverse and target myriad harms associated with deepfakes. Rather than surveying the field, this Note explores solutions to a particular set of concerns—those related to national security and election integrity. These interests deal less with personal injury and more with threats to our social order, where a bad actor could use deepfakes to exploit social divisions, undermine democratic discourse, and erode trust in governmental institutions. The resulting tangible harms could have wide-reaching effects on election integrity, military operations, and intelligence-gathering missions, among other things. This Note details harms related to election interference and national security, explores existing and proposed legislation aimed at regulating deepfakes, and examines the legal and constitutional constraints on any such legislation. Finally, this Note distills and assesses key facets of existing proposals and attempts to craft an effective legislative solution that accommodates, anticipates, and mitigates such harms without chilling expression or technological progress.

[*] J.D. candidate, Boston University School of Law, 2021; B.S., Civil Engineering, Iowa State University, 2018. I would like to thank Professor Stacey Dogan for her insight and guidance throughout the entire writing process and Kaela Dunn, Ryan Giggi, and the *Boston University Law Review* staff and editors for their outstanding editorial work. I would also like to thank Professor Marni Goldstein Caputo for teaching me legal writing after I spent four years studying concrete and bridges. Finally, I would like to thank my family and friends, Lizzy Langa and Bri Allan in particular, for their help and support.

CONTENTS

INTRODUCTION

Mark Zuckerberg sits at a desk, speaking into a camera for what appears to be a news segment. Chyrons state that Facebook is "increasing transparency on ads" as Zuckerberg announces new measures to protect elections[1]: "Imagine this for a second: one man with total control of billions of people's stolen data. All their secrets, their lives, their futures. I owe it all to Spectre. Spectre showed me that whoever controls the data controls the future."[2]

Kim Kardashian West leans against a high-top table, discussing her social media empire:

> When there's so many haters, I really don't care because their data has made me rich beyond my wildest dreams. My decision to believe in Spectre literally gave me my ratings and my fanbase. I feel really blessed because I genuinely love the process of manipulating people online for money.[3]

Although the videos of Zuckerberg and Kardashian West appear real, with lifelike images and believable (albeit suspiciously candid) quotes, they are inauthentic digital creations. Artists Bill Posters and Daniel Howe collaborated with artificial intelligence ("AI") technology startups to create the Zuckerberg and Kardashian West videos for Spectre, an art installation aiming to reveal how technology companies, advertising firms, and political campaigns influence and predict voter behavior "online and in the voting booth."[4] The videos, known as deepfakes, a portmanteau of "deep learning" and "fake," were created using a "proprietary AI algorithm, trained on 20 to 45 second scenes of the target face for between 12-24 hours."[5] The algorithm examined selected video clips, including "videos of the voice actor speaking, and then reconstructed the frames [of the] video to match the facial movements of the voice actor."[6] Deepfake technology has grown more sophisticated since these deepfakes debuted in May 2019, allowing for increasingly lifelike digital representations.

In other words, a deepfake is a highly realistic digital image, often a video, created with AI.[7] Deepfakes include face swaps, audio clips, facial reenactment,

[1] Bill Posters (@bill_posters_uk), INSTAGRAM (June 7, 2019), https://www.instagram.com/p/ByaVigGFP2U/.

[2] *Id.*

[3] Bill Posters (@bill_posters_uk), INSTAGRAM (June 1, 2019), https://www.instagram.com/p/ByKg-uKlP4C/.

[4] Press Release, Bill Posters, Gallery: 'Spectre' Launches (May 29, 2019), http://billposters.ch/spectre-launch/ [https://perma.cc/DRZ3-NDMV].

[5] Samantha Cole, *This Deepfake of Mark Zuckerberg Tests Facebook's Fake Video Policies*, VICE: MOTHERBOARD (June 11, 2019, 2:25 PM), https://www.vice.com/en_us/article/ywyxex/deepfake-of-mark-zuckerberg-facebook-fake-video-policy [https://perma.cc/3NLY-C2KU].

[6] *Id.*

[7] James Vincent, *Why We Need a Better Definition of 'Deepfake*,' VERGE (May 22, 2018, 2:53 PM), https://www.theverge.com/2018/5/22/17380306/deepfake-definition-ai-

and lip-synching.[8] Until recently, fake videos were scarce and mostly unrealistic due to "the lack of sophisticated editing tools, the high demand on domain expertise, and the complex and time-consuming process involved."[9] Now, deepfakes are commonplace and easily created because of "the accessibility to large-volume training data and high-throughput computing power" and "the growth of machine learning and computer vision techniques that eliminate the need for manual editing steps."[10]

At the most basic level, a deepfake is created by taking an input source of videos or other images of an individual and outputting a manipulated video of that individual's face.[11] The output is generated by neural networks, which are "a means of doing machine learning, in which a computer learns to perform some task by analyzing training examples."[12] The neural network is trained to "automatically map the facial expressions of the source" to create the manipulated video.[13]

The newest technology to create deepfakes comes from generative adversarial networks ("GANs"), which "consist[] of two deep neural networks trained in tandem."[14] One neural network, "known as the 'actor,' tries to learn the statistical patterns in a data set, such as a set of images or videos, and then generate convincing synthetic pieces of data."[15] The second neural network, "called the 'critic,' tries to distinguish between real and fake examples."[16] The result is an iterative process in which the feedback from the second neural network enables the first neural network to produce increasingly realistic deepfakes.[17] In essence, the first neural network is a picture forger; the second neural network is an art detective.[18] The two go back and forth, trying to outwit

manipulation-fake-news.

[8] *Id.*

[9] Yuezun Li & Siwei Lyu, Exposing DeepFake Videos by Detecting Face Warping Artifacts 1 (May 22, 2019) (unpublished manuscript), https://arxiv.org/pdf/1811.00656.pdf [https://perma.cc/35BZ-U694].

[10] *Id.*

[11] *See id.* at 2.

[12] Larry Hardesty, *Explained: Neural Networks*, MIT NEWS (Apr. 14, 2017), http://news.mit.edu/2017/explained-neural-networks-deep-learning-0414 [https://perma.cc/593N-6MWX].

[13] Li & Lyu, *supra* note 9, at 1.

[14] *Id.* at 2.

[15] Will Knight, *The US Military Is Funding an Effort to Catch Deepfakes and Other AI Trickery*, MIT TECH. REV. (May 23, 2018), https://www.technologyreview.com/2018/05/23/142770/the-us-military-is-funding-an-effort-to-catch-deepfakes-and-other-ai-trickery/.

[16] *Id.*

[17] *Id.*

[18] Martin Giles, *The GANfather: The Man Who's Given Machines the Gift of Imagination*, MIT TECH. REV. (Feb. 21, 2018), https://www.technologyreview.com/2018/02/21/145289/the-ganfather-the-man-whos-given-machines-the-gift-of-imagination/.

each other until the art detective can no longer tell what is real and what is not.[19] Similar technology is used to create deepfake audio clips.[20]

This technology is cheap and becoming increasingly accessible to amateurs.[21] In September 2019, deepfake pioneer Hao Li predicted that "'perfectly real' [deepfakes] will be accessible to everyday people" within six months to a year.[22] Samsung has already developed commercially available software that can create a highly realistic deepfake with a single image, using a neural network trained on a large data set of photos and videos.[23] Because machine learning is often publicly available through various commercial services,[24] the capacity to easily

[19] *Id.*

[20] For example, in 2017, Lyrebird AI developed "voice-cloning" technology and released fake audio clips, including a clip of President Donald Trump discussing sanctions against North Korea. Lyrebird AI (@LyrebirdAi), TWITTER (Sept. 4, 2017, 2:41 AM), https://twitter.com/LyrebirdAi/status/904595326929174528 (demonstrating such technology with audio clip of President Trump appearing to say "[t]he United States is considering, in addition to other options, stopping all trade with any country doing business with North Korea"); *see also* James Vincent, *Lyrebird Claims It Can Recreate Any Voice Using Just One Minute of Sample Audio*, VERGE (Apr. 24, 2017, 12:04 PM), https://www.theverge.com/2017/4/24/15406882/ai-voice-synthesis-copy-human-speech-lyrebird (finding Lyrebird AI's technology "impressive" and predicting that it would "no doubt improve over time," but remaining skeptical of its overall efficacy). Fake audio clips have already been used to defraud a company, which suggests that the authenticity of fake audio clips may be more difficult to discern than video and therefore pose a more immediate threat. *See* Catherine Stupp, *Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case*, WALL ST. J. (Aug. 30, 2019, 12:52 PM), https://www.wsj.com /articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 (recounting March 2019 episode in which the chief executive officer ("CEO") of a U.K.-based energy company was convinced to transfer €220,000 by phone call convincingly replicating the voice of the CEO's boss).

[21] Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1762-63 (2019).

[22] Kevin Stankiewicz, *'Perfectly Real' Deepfakes Will Arrive in 6 Months to a Year, Technology Pioneer Hao Li Says*, CNBC: TECH (Jan. 17, 2020, 2:51 AM), https://www.cnbc.com/2019/09/20/hao-li-perfectly-real-deepfakes-will-arrive-in-6-months-to-a-year.html [https://perma.cc/K6DJ-NMC6].

[23] Joan E. Solsman, *Samsung Deepfake AI Could Fabricate a Video of You from a Single Profile Pic*, CNET (May 24, 2019, 7:00 AM), https://www.cnet.com/news/samsung-ai-deepfake-can-fabricate-a-video-of-you-from-a-single-photo-mona-lisa-cheapfake-dumbfake/ [https://perma.cc/B34W-7UJF].

[24] Certain businesses, such as Generated.Photos or ThisPersonDoesNotExist.com, sell A.I.-generated images of fake persons "for characters in a video game, or to make your company website appear more diverse." Kashmir Hill & Jeremy White, *Designed to Deceive: Do These People Look Real to You?*, N.Y. TIMES (Nov. 21, 2020) (citation omitted), https://www.nytimes.com/interactive/2020/11/21/science/artificial-intelligence-fake-people-faces.html. Companies like Rosebud.AI go a step further by generating deepfake images and videos. *Id.*

create deepfakes ensures that they will spread throughout the general public.[25] Compounding the problem, social media networks and online platforms provide an environment for deepfakes to circulate widely. The ability to distribute images, audio, and video cheaply and instantaneously, coupled with cognitive biases in which people blindly share negative, novel, or belief-conforming information, suggests that scandalous or harmful deepfake content will spread like wildfire through social media and online platforms.[26]

Increasingly realistic deepfakes pose a significant threat to a bevy of individual and societal interests.[27] Indeed, a realistic and widely disseminated deepfake could distort democratic discourse, manipulate elections, or jeopardize national security.[28] Members of Congress, such as Representative Yvette Clarke, have suggested that while deepfakes pose numerous threats to society, "the threat of election interference is perhaps the most menacing and urgent."[29] The

[25] *See* Chesney & Citron, *supra* note 21, at 1763.

[26] *See id.* at 1768 ("Information cascades, natural attraction to negative and novel information, and filter bubbles provide an all-too-welcoming environment as deep-fake capacities mature and proliferate.").

[27] *See id.* at 1771-86 (identifying harmful uses of deepfakes, including individual exploitation and sabotage; distorting democratic discourse; manipulating elections; eroding public trust; exacerbating social divisions; undermining public safety, diplomacy, and journalism; jeopardizing national security; and facilitating spread of fake news). Perhaps the most significant threat currently posed by deepfakes is deepfake pornography targeting women, which accounts for 96% of all deepfake videos online. *See* Aja Romano, *Deepfakes Are a Real Political Threat. For Now, Though, They're Mainly Used to Degrade Women*, VOX (Oct. 7, 2019, 7:00 PM), https://www.vox.com/2019/10/7/20902215/deepfakes-usage-youtube-2019-deeptrace-research-report. This said, it is important to note that deepfakes have positive uses as well, such as promoting artwork, education, and self-expression. *See* Chesney & Citron, *supra* note 21, at 1769-71; *see also* Jessica Silbey & Woodrow Hartzog, *The Upside of Deep Fakes*, 78 MD. L. REV. 960, 960-66 (2019) (describing "beneficial uses of deep-fake technology," not only for art and science but also because "they might help muster the political will to address the larger, structural problems made worse by the inability to trust what we see and hear").

[28] Chesney & Citron, *supra* note 21, at 1777-79, 1783-84 (reviewing instances of these harms, such as the 2017 presidential election in France where "Russians mounted a covert-action program that blended cyber-espionage and information manipulation").

[29] Yvette Clarke, *Deepfakes Will Influence the 2020 Election—and Our Economy, and Our Prison System*, QUARTZ: IDEAS (July 11, 2019), https://qz.com/1660737/deepfakes-will-influence-the-2020-election/. Senator Marco Rubio has echoed Congresswoman Clarke's concerns, stating that he believes that

> the next wave of attacks against America and Western democracies . . . is the ability to produce fake videos that can only be determined to be fake after extensive analytical analysis, and by then the election is over and millions of Americans have seen an image that they want to believe anyway because of their preconceived bias . . . .

Nomination of William R. Evanina to Be the Director of the National Counterintelligence and Security Center Before the S. Select Comm. on Intel., 115th Cong. 12 (2018) (statement of Sen. Marco Rubio, Member, S. Select Comm. on Intel.).

well-timed release of a deepfake portraying a candidate or public official taking bribes, using racial epithets, committing adultery, or, alternatively, engaging in positive behavior could change the outcome of an election or erode public trust in governmental institutions.[30] While false information and conspiracy theories are nothing new, weaponized deepfakes present additional and more complicated challenges in combatting threats posed by misinformation.[31] Perhaps more troubling, deepfakes have the capacity to provoke armed conflict, which poses a unique threat to foreign and domestic national security. Such a conflict occurred in 2019 when a purported deepfake of Gabonese President Ali Bongo Ondimba created uncertainty around his health and helped precipitate an attempted coup d'état.[32] The video's authenticity even baffled experts, who cited irregular facial movements and speech patterns as signs of a deepfake but could not ultimately determine whether the video was real.[33] Incidents like this illustrate another way that deepfakes can wreak havoc: the mere knowledge that deepfakes exist can plant a seed of doubt about any video's authenticity in the mind of a viewer. Whether a video is ultimately revealed to be real or fake does not matter when the consequences of its dissemination are irreparable.

Although deepfakes have obvious potential harms, they also have the capacity for unambiguously beneficial uses in education, artwork, and promotion of democratic institutions.[34] For example, deepfakes can facilitate personal

[30] *See* Chesney & Citron, *supra* note 21, at 1778. Indeed, both positive and negative deepfakes of prominent public officials have already surfaced. For example, in 2019, a French charity created a deepfake video of President Trump to raise awareness of their efforts to eradicate AIDS. *See* Solidarité Sida (@SolidariteSida), TWITTER (Oct. 7, 2019, 2:00 AM), https://twitter.com/SolidariteSida/status/1181086753693810689 (showing what appears to be President Trump saying, "Today is a historic day. I have tremendous news. Today we eradicated AIDS"). The video was only used as part of an advertisement campaign and included a disclaimer, but the capacity to spread misinformation about global health crises raises similar concerns. Indeed, although not disseminated via deepfakes, misinformation on social media about false origins of COVID-19 and sham cures for the disease likely contributed to its early spread in the United States. *See* Tony Romm, *Social Media Sites Scramble to Stop Spread of Misinformation, Falsehoods*, WASH. POST, Jan. 28, 2020, at A13; Peter Suciu, *During COVID-19 Pandemic It Isn't Just Fake News but Seriously Bad Misinformation that Is Spreading on Social Media*, FORBES (Apr. 8, 2020, 7:00 AM), https://www.forbes.com/sites/petersuciu/2020/04/08/during-covid-19-pandemic-it-isnt-just-fake-news-but-seriously-bad-misinformation-that-is-spreading-on-social-media/#22f999797e55 [https://perma.cc/5HMS-Z4YV].

[31] These challenges center on the inherent believability of a well-manufactured deepfake. *See, e.g.*, *supra* notes 20, 27 and 30.

[32] Ali Breland, *The Bizarre and Terrifying Case of the "Deepfake" Video that Helped Bring an African Nation to the Brink*, MOTHER JONES (Mar. 15, 2019), https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/ [https://perma.cc/9X8Z-ELCG].

[33] *Id.*

[34] Silbey & Hartzog, *supra* note 27, at 960-66; *see also supra* note 27.

expression and autonomy, as demonstrated by the text-to-speech technology company CereProc, which created "a digital voice for a radio host who lost his voice due to a medical condition."[35] Deepfakes have also been used to recreate historical figures[36] and broadcast educational campaigns to combat the spread of diseases.[37] In addition, deepfakes have been used to depict deceased actors in films, parody public officials, or create expressive illustrations that are otherwise unavailable.[38] Some commentators have also argued that the uncertainty created by the mere existence of deepfakes is actually beneficial, as media companies and democratic institutions are forced to strengthen and become more transparent in response to potential harms posed by deepfakes.[39] Thus, an outright ban on deepfakes would be ill-advised.

Given the rapid development of deepfake technology, there is a growing notion that the dangers they pose justify regulation at the state and federal levels.[40] Proposed regulation comes in many forms and targets myriad harms associated with deepfakes.[41] Rather than surveying the field, this Note focuses on solutions to concerns related to national security and election integrity. Here, our social order is at stake—a bad actor could use deepfakes to exploit social divisions, undermine democratic discourse, and erode trust in governmental

[35] William Turton & Andrew Martin, *How Deepfakes Make Disinformation More Real Than Ever*, BLOOMBERG (Jan. 7, 2020, 4:30 AM), https://www.bloomberg.com/news/articles/2020-01-06/how-deepfakes-make-disinformation-more-real-than-ever-quicktake. Deepfakes could also be used to facilitate self-expression by allowing "individuals suffering from certain physical disabilities [to] interpose their faces and that of consenting partners into pornographic videos, enabling virtual engagement with an aspect of life unavailable to them in a conventional sense." Chesney & Citron, *supra* note 21, at 1771.

[36] CereProc recreated a version of an undelivered speech by President John F. Kennedy, which he was scheduled to give on the day he was assassinated. Turton & Martin, *supra* note 35.

[37] In an effort to stop the spread of malaria, the software and marketing firm Synthesia created a deepfake video of former soccer star David Beckham speaking in nine languages on behalf of Malaria Must Die. Malaria Must Die, *David Beckham Speaks Nine Languages to Launch Malaria Must Die Voice Petition*, YOUTUBE (Apr. 9, 2019), https://www.youtube.com/watch?v=QiiSAvKJIHo.

[38] *See* Chesney & Citron, *supra* note 21, at 1770.

[39] Silbey & Hartzog, *supra* note 27, at 961 (describing how "education, journalism, and representative democracy . . . could be strengthened as a response to deep fakes").

[40] *See infra* Part I (demonstrating adverse consequences of malicious deepfakes through examples of national security and election interference). *But see* Hayley Tsukayama, India McKinney & Jamie Williams, *Congress Should Not Rush to Regulate Deepfakes*, ELEC. FRONTIER FOUND. (June 24, 2019), https://www.eff.org/deeplinks/2019/06/congress-should-not-rush-regulate-deepfakes [https://perma.cc/EC94-798N] (arguing that companies should be allowed to self-moderate without hastily written antideepfake statutes that "lead[] to de facto government regulation of speech").

[41] *See infra* Part II (discussing "patchwork" of legislation either directly targeting or indirectly applicable to deepfakes targeting national security and election integrity).

institutions. The resulting harms could have wide-reaching effects on election integrity, military operations, and intelligence gathering, among other things.[42]

Deepfakes do have many beneficial uses, making an outright ban imprudent. However, the threats that deepfakes pose demand a uniform and effective law that punishes harmful use of deepfakes and deters prospective bad actors. A comprehensive federal legislative solution is desirable because deepfakes can spread quickly, reaching millions of people across dozens of jurisdictions.[43] A consistent federal response thus offers a more predictable and effective solution. Laws in Texas and California, as well as proposed federal legislation, provide a helpful blueprint for an effective federal law. Such legislation must effectively and broadly define what a deepfake is, anticipate technological changes, and comport with established national security and First Amendment law and precedent.

Part I of this Note explores harms that deepfakes pose to national security and election integrity. Part II surveys the emerging patchwork of state and federal legislation concerning various aspects of deepfakes. Although not all existing legislation relates directly to national security and election integrity, key principles informing a comprehensive solution can be drawn from these regimes. Part III considers how national security and First Amendment precedents will limit the scope of federal legislation and whether a shifting technological environment warrants altering traditional doctrinal approaches. Part IV analyzes the existing legislation directed at election interference and threats to national security, considers established legal and constitutional constraints to developing deepfake legislation, and proposes legislation that synthesizes effective and desirable facets of existing laws and proposals.

I. DEEPFAKE HARMS

To effectively combat threats posed by deepfakes, it is first necessary to identify and understand the underlying harms and their potential consequences. Such considerations should include input from industry leaders in science, technology, and government.[44] Identifying underlying harms and the ways in which deepfakes influence their viewers is the first step in crafting a targeted legislative solution.

[42] *See* Chesney & Citron, *supra* note 21, at 1777.

[43] *See* Cass R. Sunstein, *Constitutional Caution*, 1996 U. CHI. LEGAL F. 361, 365 ("It may well be that the easy transmission of such material to millions of people will justify deference to reasonable legislative judgements.").

[44] Clarke, *supra* note 29 (asserting that "a comprehensive non-partisan industry-expert partnership system" is necessary to prevent harms that mere detection systems will not undo or prevent).

A. *Threats to National Security*

In the hands of bad actors, deepfakes pose significant threats to our national security. Lawmakers have expressed concerns about the implications of the quick spread of technology enabling the creation of deepfakes, calling on the Director of National Intelligence to assess threats to national security and arguing that such technology could be used to spread misinformation, exploit social division, and create political unrest.[45]

Misinformation spread through deepfakes could jeopardize national security in myriad ways. For example, misinformation could jeopardize the safety of military forces engaging with a foreign civilian population if a deepfake circulated depicting military members disparaging, assaulting, or killing civilians.[46] A bad actor could take advantage of a region's instability by using a deepfake to inflame a local population, which could lead to civilian casualties, greater enemy recruitment, or violent confrontations with U.S. personnel.[47] Hostile foreign regimes could also use deepfakes to create propaganda, depicting world leaders "shouting offensive phrases or ordering atrocities."[48] Highly realistic deepfakes thus pose a unique threat to public safety and national security because of their persuasive power, bolstered by "the distribution powers of social media."[49]

Deepfakes could also pose a threat to national security if they are used to engage in wartime deception.[50] In this context, bad actors could use deepfakes to impersonate military or intelligence officers "ordering the sharing of sensitive

[45] More specifically, members of congress have argued,

> Forged videos, images or audio could be used . . . by foreign or domestic actors to spread misinformation. As deep fake technology becomes more advanced and more accessible, it could pose a threat to United States public discourse and national security, with broad and concerning implications for offensive active measures campaigns targeting the United States.
>
> Given the significant implications of these technologies and their rapid advancement, we believe that a thorough review by the Intelligence Community is appropriate, including an assessment of possible counter-measures and recommendations to Congress.

Letter from Adam B. Schiff, Stephanie Murphy & Carlos Curbelo, Members, U.S. Cong., to Daniel R. Coats, Dir. of Nat'l Intel. (Sept. 13, 2018), https://schiff.house.gov/imo/media/doc/2018-09%20ODNI%20Deep%20Fakes%20letter.pdf [https://perma.cc/DS92-TZN4].

[46] *See* Chesney & Citron, *supra* note 21, at 1783.

[47] *See id.*

[48] GREG ALLEN & TANIEL CHAN, HARVARD KENNEDY SCH. BELFER CTR. FOR SCI. & INT'L AFFS., ARTIFICIAL INTELLIGENCE AND NATIONAL SECURITY 32 (2017), https://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf [https://perma.cc/F3EW-3367].

[49] *See* Chesney & Citron, *supra* note 21, at 1781.

[50] *See id.* at 1783.

information or taking some action that would expose forces to vulnerability."[51] Combined with a cyberattack, such as a hack of a news organization's website or of a trove of government documents, a deepfake could be widely disseminated and threaten an entire governmental regime or national economy.[52] For example, an adversary could acquire confidential documents and selectively leak deepfake forgeries along with the real documents.[53] Once real or forged documents are released, governmental officials "would also face major difficulty in limiting and remediating the potentially significant consequences of [a] false understanding."[54]

Of course, the most catastrophic scenario stemming from a convincing deepfake is nuclear war brought on by a forged video of a world leader declaring war or threatening retaliation.[55] While such drastic consequences may seem unrealistic, international nuclear posturing over false information disseminated through the Internet has happened before. In 2016, the Pakistani defense minister wrote a threatening tweet directed at Israel after reading a false report that purported to show Israel threatening Pakistan with nuclear weapons.[56]

Like with other uses of deepfakes, however, these harms are not without benefits. In the military and national security context, the same AI technology used to create deepfakes could be repurposed to assist intelligence agencies in surveilling hostile areas by detecting human threats with facial recognition software, thereby countering guerilla warfare and insurgency.[57] The U.S. military could also use AI technology to achieve a strategic advantage over enemies with lesser capabilities.[58] Clearly, completely curtailing the development of AI technology in an effort to combat deepfakes would be disadvantageous. A tailored solution that specifically addresses tangible national

[51] ALLEN & CHAN, *supra* note 48, at 33. This threat seems particularly pressing, considering the demonstrated ability of deepfakes to defraud companies over the phone. Stupp, *supra* note 20 (recounting story of a CEO defrauded by deepfake voice of their boss on the telephone).

[52] *See* ALLEN & CHAN, *supra* note 48, at 33-34.

[53] *Id.* at 34. Prohibiting the release of confidential government information along with selective deepfakes raises First Amendment concerns, as discussed in Part II.

[54] *Id.*

[55] Hany Farid, professor of computer science at Dartmouth University, raised this exact hypothetical at a 2018 Defense Advanced Research Projects Agency Media Forensics program meeting. Jon Christian, *Experts Fear Face Swapping Tech Could Start an International Showdown*, OUTLINE (Feb. 1, 2018, 11:38 AM), https://theoutline.com/post/3179/deepfake-videos-are-freaking-experts-out?zd=1&zi=adwon5jm [https://perma.cc/EHJ5-68LW].

[56] Russell Goldman, *Reading Fake News, Pakistani Minister Directs Nuclear Threat at Israel*, N.Y. TIMES (Dec. 24, 2016), https://www.nytimes.com/2016/12/24/world/asia/pakistan-israel-khawaja-asif-fake-news-nuclear.html.

[57] *See* ALLEN & CHAN, *supra* note 48, at 31-32.

[58] *See id.* at 32 (arguing that countries with more advanced AI technology will have a comparative military advantage).

security and electoral threats stemming from deepfakes while leaving breathing room for innovation is desirable.

B. *Election Interference*

The release of a well-timed deepfake also has the potential to interfere in state and federal elections by injecting convincing falsehoods and uncertainty concerning candidates' personal lives and policy positions into the electoral process. Such uncertainty could undermine faith in the outcome of that election.[59] Acknowledging these risks, the former Director of National Intelligence has identified deepfakes as a threat that adversaries could use against the United States and its allies by manipulating or disrupting their election systems.[60] Because deepfake technology is broadly accessible, any actors—state sponsored or not—could generate and disseminate deepfakes targeted at disrupting elections.[61] While it is doubtful that malicious deepfakes have meaningfully harmed U.S. elections thus far, the technology continues to advance rapidly, necessitating a solution before this harm is fully realized.[62]

[59] *See Worldwide Threat Assessment of the U.S. Intelligence Community, Hearing Before the S. Select Comm. on Intel.*, 116th Cong. 7 (2019) [hereinafter Coats Statement] (prepared statement of Daniel R. Coats, Dir. of Nat'l Intel.) ("Adversaries and strategic competitors . . . may seek to use cyber means to directly manipulate or disrupt election systems—such as by tampering with voter registration or disrupting the vote tallying process—either to alter data or to call into question our voting process."); Chesney & Citron, *supra* note 21, at 1778.

[60] Coats Statement, *supra* note 59, at 7.

[61] Chesney & Citron, *supra* note 21, at 1779.

[62] Malicious deepfakes do not appear to have had any meaningful effect on the 2020 election. Two notable deepfake campaigns were created to warn users about the potential harms flowing from the misuse of deepfake technology. Tom Simonite, *What Happened to the Deepfake Threat to the Election?*, WIRED (Nov. 16, 2020, 7:00 AM), https://www.wired.com/story/what-happened-deepfake-threat-election/. Democratic House candidate Phil Ehr deployed a deepfake during the election in a campaign ad that featured his opponent, Republican Congressman Matt Gaetz, making claims such as "Obama is way cooler than me." Ehr for Congress, *#DeepFake*, YOUTUBE (Oct. 1, 2020), https://www.youtube.com/watch?v=Y6HKo-IAltA. However, Ehr himself appeared at the end of the ad to explicitly inform viewers that the depiction of Gaetz was a deepfake. *Id.* Other deepfake ads warned about the malicious use of deepfakes, including deepfake videos from RepresentUs that depicted Vladimir Putin and Kim Jong-Un discussing the erosion of democracy but were explicitly labeled as inauthentic. RepresentUs, *Dictators - Vladimir Putin*, YOUTUBE (Sept. 29, 2020), https://www.youtube.com/watch?v=sbFHhpYU15w&feature=youtu.be; RepresentUs, *Dictators - Kim Jong-Un*, YOUTUBE (Sept. 29, 2020), https://www.youtube.com/watch?v=ERQlaJ_czHU&feature=youtu.be.

One notable use of a malicious deepfake likely intended to interfere with the 2020 presidential election promoted conspiracy theories aimed at Hunter Biden, the son of President Biden. *E.g.*, Ben Collins & Brandy Zadrozny, *How a Fake Persona Laid the Groundwork for a Hunter Biden Conspiracy Deluge*, NBC News (Oct. 30, 2020, 11:19 AM), https://www.nbcnews.com/news/amp/ncna1245387 [https://perma.cc/G97P-YHAW]. The

A well-timed deepfake, distributed when there is "enough window for the fake to circulate but not enough window for the victim to debunk it effectively," could influence the outcome of an election by creating a "decisional chokepoint[]: [a] narrow window[] of time during which irrevocable decisions are made, and during which the circulation of false information therefore may have irremediable effects."[63] In this manner, deliberately distorted videos amplified through social media could "cloud reality at a time when the existence of objective facts increasingly has been called into question."[64] Along these lines, some commentators have suggested that the highest priority of social media platforms should be to remove provably false content, such as deepfakes, that affect democratic institutions.[65]

The debate over election harms posed by deepfakes is itself a threat, as it creates uncertainty and undermines faith in the reliability of video images found online. If a compromising video or audio clip was released, a candidate could attack its credibility, despite knowing of its legitimacy.[66] In this way, public

debunked conspiracy theory was based in part on a fake intelligence document authored by a security analyst purportedly named Martin Aspen. *Id.* However, researchers discovered that Aspen was a fabricated identity and "that Aspen's profile picture was created with an artificial intelligence face generator." *Id.* Such use of a deepfake illustrates how an effective solution targeting the harms posed by deepfakes must address deepfakes' intentions to harm, persuasiveness, and believability. The deepfake profile picture initially lent credibility to the document because it was believable; it appeared to depict a real person and could not be easily debunked as a fabricated profile because the photo was not, for example, a stock photo found on the Internet. This implied that, at the very least, someone with security credentials had researched and compiled the document. The photo's believability in turn likely bolstered the document's persuasiveness before it was widely debunked. *See id.* Persuasiveness and believability, coupled with the intent to interfere in an election, is the recipe for tangible harm caused by deepfakes.

[63] Chesney & Citron, *supra* note 21, at 1778.

[64] PAUL M. BARRETT, NYU STERN CTR. FOR BUS. & HUM. RTS., DISINFORMATION AND THE 2020 ELECTION: HOW THE SOCIAL MEDIA INDUSTRY SHOULD PREPARE 3-4 (2019).

[65] *Id.* at 4. Indeed, in early 2020, Facebook announced that they would ban deepfakes and other misleading or manipulated media that meets certain criteria:

> - It has been edited or synthesized – beyond adjustments for clarity or quality – in ways that aren't apparent to an average person and would likely mislead someone into thinking that a subject of the video said words that they did not actually say. And:
> - It is the product of artificial intelligence or machine learning that merges, replaces or superimposes content onto a video, making it appear to be authentic.
>
> This policy does not extend to content that is parody or satire, or video that has been edited solely to omit or change the order of words.

Monika Bickert, *Enforcing Against Manipulated Media*, FACEBOOK (Jan. 6, 2020), https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/ [https://perma.cc/8JAQ-WB3J].

[66] *See, e.g.*, Tom Simonite, *Will 'Deepfakes' Disrupt the Midterm Election?*, WIRED (Nov. 1, 2018, 7:00 AM), https://www.wired.com/story/will-deepfakes-disrupt-the-midterm-election/ [https://perma.cc/NE9D-2DNX] ("The biggest tangible threat of deepfakes so far is

officials could deny the truth of an image, video, or audio clip, as a "skeptical public will be primed to doubt the authenticity of real audio and video evidence."[67] Such distrust in the authenticity of images, videos, and audio clips of elected officials further erodes trust in the electoral process and governmental institutions as a whole.[68]

In sum, deepfakes pose significant current and future threats to U.S. national security and electoral processes. Advancing deepfake technology will only exacerbate such harms, necessitating a targeted legislative solution.

II. Existing Patchwork of State and Federal Legislation

Given the nascent state of deepfake technology, the legislative and legal regime governing deepfakes has yet to fully develop. While some scholars have suggested using existing legal regimes to bring civil and criminal actions,[69] lawmakers did not consider the technology and effects of deepfakes when drafting such legislation. Presently, a handful of states serve as laboratories,[70] combating different threats posed by deepfakes by creating individual rights of action or protecting against election interference, among other methods. Given the rapid and unpredictable advancement of deepfakes, however, we cannot afford to wait for the results of these state-based experiments. While various states have passed legislation aimed at preventing election interference by deepfakes, federal legislation is necessary to best address the problem. From an administrative efficiency perspective, a federal solution allows for a more effective and centralized response than a patchwork of state laws and

the allegation that any future hot mike or covert recording of Donald Trump or any other candidate would be a deepfake . . . .").

[67] Chesney & Citron, *supra* note 21, at 1785. Politicians have already attempted to dismiss videos by questioning their authenticity. In 2019, a video of a man allegedly confessing to having sex with a Malaysian cabinet minister was questioned as a deepfake. Drew Harwell, *Scramble Is On to Detect, Stop 'Deepfake' Videos*, WASH. POST, June 13, 2019, at A1.

[68] Chesney & Citron, *supra* note 21, at 1779.

[69] *See* Russell Spivak, *"Deepfakes": The Newest Way to Commit One of the Oldest Crimes*, 3 GEO. L. TECH. REV. 339, 364 (2019) ("In light of the potential havoc deepfakes and related technologies can wreak, scholars and legislators alike ought to consider how to structure relevant legal regimes. The constitutionality of proactive legislation is dubious . . . . Thus, the focus on rectifying harms to victims should explore other methods . . . ."). Spivak argues that tort actions, including defamation, various privacy torts, and rights of publicity, can be brought under state law. *Id.* at 364-86. Chesney and Citron offer a more expansive list: various tort actions as civil remedies; criminal liability for cyberstalking, impersonation, fraud, or incitement; administrative agency actions brought by the FCC, FEC, or FTC; coercive responses including military actions or sanctions; and market solutions wherein private companies identify and police deepfakes on the Internet. Chesney & Citron, *supra* note 21, at 1786-819.

[70] *See* New State Ice Co. v. Liebmann, 285 U.S. 262, 311 (1932) (Brandeis, J., dissenting) (noting that states are in the best position to experiment with novel laws that may not be ready for federal application).

enforcement mechanisms.[71] Relying on an assortment of state laws could help wrongdoers avoid detection or exploit differences between jurisdictions. Moreover, a uniform federal solution best provides "concentrated action" to anticipate and remedy disparate harms posed by deepfakes.[72] Several bills regulating deepfakes have been put forth in Congress, but only one directly addresses deepfakes and election interference.[73] This Part proceeds by surveying the body of existing and proposed state and federal law.

A. *State Legislation Aimed at Election Interference*

To date, Texas and California are the only states to propose or enact legislation prohibiting deepfakes for the purpose of preventing election interference.[74] Although both serve as helpful guideposts in analyzing the existing landscape governing deepfakes, the California legislation is substantially more detailed in defining what constitutes deepfake technology and in creating punishable offenses and exceptions.[75]

Texas. Texas was the first state to prohibit the creation of deepfakes designed to interfere with elections.[76] Texas Senate Bill 751 ("Texas Deepfake Act") prohibits a person from intending to "injure a candidate or influence the result of an election" by creating a deepfake and causing it "to be published or distributed within 30 days of an election."[77] The Act defines a deepfake as a "video, created with the intent to deceive, that appears to depict a real person performing an action that did not occur in reality."[78] Violation of the Texas

[71] *See* Taleed El-Sabawi, *MHPAEA & Marble Cake: Parity & the Forgotten Frame of Federalism*, 124 DICK. L. REV. 591, 604 (2020).

[72] *Id.* ("While states' capacity to administer public policy has increased, when broad policy goals require concentrated action, federal dominance is preferable . . . ." (footnote omitted)).

[73] DEEP FAKES Accountability Act, H.R. 3230, 116th Cong. (2019); *see also* MATTHEW F. FERRARO, WILMERHALE, DEEPFAKE LEGISLATION: A NATIONWIDE SURVEY—STATE AND FEDERAL LAWMAKERS CONSIDER LEGISLATION TO REGULATE MANIPULATED MEDIA § I (2019) ("[T]he DEEP FAKES Accountability Act seeks to protect against the full gamut of deepfake harms, from nonconsensual pornography to foreign interference in elections and public policy debates, from inciting violence to conducting financial fraud and identity theft.").

[74] *See* TEX. ELEC. CODE ANN. § 255.004 (West 2019); CAL. ELEC. CODE § 20010 (West 2020); Assemb. B. 1280, 2019-2020 Leg., Reg. Sess. (Cal. 2019); *see also* FERRARO, *supra* note 73, § V (describing antideepfake legislation in Texas).

[75] CAL. ELEC. CODE § 20010 (West 2020); Cal. Assemb. B. 1280. Once the law definitively holds wrongdoers civilly or criminally liable, the approach to solving the national security and election integrity problems posed by deepfakes will likely warrant additional considerations and changes. However, at the time of writing, I am unaware of any final legal outcome under any of the deepfake laws presented in this Part.

[76] TEX. ELEC. CODE ANN. § 255.004 (West 2019).

[77] S.B. 751, 86th Leg., Reg. Sess. (Tex. 2019) (enacted).

[78] TEX. ELEC. CODE ANN. § 255.004(e) (West 2019).

Deepfake Act is a Class A misdemeanor punishable with up to a year in county jail and a $4,000 fine.[79]

California. The California legislature has passed one bill concerning deepfakes and election interference and considered a second. The enacted bill, titled the "Elections: Deceptive Audio or Visual Media Act" and codified in section 20010 of the California Election Code, prohibits a

> person, committee . . . or other entity, . . . within 60 days of an election at which a candidate for elective office will appear on the ballot, [from] distribut[ing], with actual malice, materially deceptive audio or visual media . . . of the candidate with the intent to injure the candidate's reputation or to deceive a voter into voting for or against the candidate . . . [unless the] media includes a disclosure stating [that the media] "has been manipulated."[80]

The statute does not explicitly refer to deepfakes but instead prohibits images, audio, and video recordings that "would falsely appear to a reasonable person to be authentic" and "would cause a reasonable person to have a fundamentally different understanding or impression of the expressive content . . . than that person would have if the person were hearing or seeing the unaltered, original version."[81] The statute provides exceptions for "materially deceptive audio or visual media" constituting satire and parody.[82]

Finally, the Act creates a private right of action. Under the Act, a candidate for elected office "whose voice or likeness appears in a materially deceptive audio or visual media distributed in violation of [the Act]" is authorized to seek equitable relief and recover damages against the distributor.[83]

The second bill, California Assembly Bill 1280, titled the "Crime: Deceptive Recordings Act," criminalizes the preparation, production, or development, "without the depicted individual's consent, [of] a deepfake with the intent that the deepfake coerce or deceive any voter into voting for or against a candidate or measure in an election that is occurring within 60 days."[84]

[79] *Id.* § 255.004(c); TEX. PENAL CODE ANN. § 12.21 (West 2019).

[80] CAL. ELEC. CODE § 20010 (West 2020).

[81] *Id.* § 20010(e)(1)-(2).

[82] *Id.* § 20010(d)(5).

[83] *Id.* § 20010(c)(1)-(2).

[84] Assemb. B. 1280, 2019-2020 Leg., Reg. Sess. (Cal. 2019). Assembly Bill 1280 failed in committee on February 3, 2020, and it was subsequently filed with the Chief Clerk of the Assembly for reconsideration. *AB-1280 Crimes: Deceptive Recordings.*, CAL. LEGIS. INFO. (Apr. 22, 2019, 9:00 PM), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB1280 [https://perma.cc/T4W2-KJXJ] (click "History"). Thereafter, it remained inactive for over a year and is considered dead at the time of writing. *Id.* (click "Status").

B. *State Legislation Aimed at Other Harms*

Several states have passed legislation aimed at preventing harms from deepfakes in other contexts. Although not directly applicable, deepfake legislation in other contexts can provide guidance in creating analogous legislation regarding election interference or national security. Bills and laws in Virginia, the focus of which is combatting deepfake pornography, and New York, the focus of which is creating rights to one's digital likeness, provide creative solutions for protecting an individual's right to privacy and insight into shaping legislation that targets the creators and distributors of malicious deepfakes. Moreover, proposed legislation in Massachusetts provides a blueprint for broadly criminalizing the use of deepfakes to commit illegal acts.

Virginia. In response to deepfake software that enabled users to "make images of clothed women appear to be realistic nudes,"[85] Virginia became the first state to criminalize the distribution of nonconsensual deepfake pornography in 2019.[86] The law prohibits an unauthorized or unlicensed person who has the intent to "coerce, harass, or intimidate" from disseminating or selling any "videographic or still image[s] created by any means whatsoever" depicting a "person who is totally nude, or in a state of undress."[87]

New York. In 2018, New York legislators introduced a bill that, if passed,[88] would have established a right of privacy and publicity to a person's "digital replica," extending the right to one's digital likeness for forty years after death.[89] The bill received significant criticism from movie studios and media companies, such as Disney, NBCUniversal, Viacom, and Warner Brothers, which warned against restricting the ability to depict real-life individuals in biopics or other films.[90] This criticism was incorporated into a new version of this bill, which creates a private right of action over the "unlawful dissemination or publication of a sexually explicit depiction of an individual."[91] It passed both houses of the

[85] FERRARO, *supra* note 73, § VI. This software, called "DeepNude," caused concern about the ease with which amateurs could create sexual imagery of others without their consent. *Id.*

[86] VA. CODE ANN. § 18.2-386.2(A) (2020).

[87] *Id.*

[88] This bill failed to pass the Senate and has not been reconsidered. *See Assembly Bill A8155B*, N.Y. ST. SENATE, https://www.nysenate.gov/legislation/bills/2017/a8155 [https://perma.cc/X2WP-W5D3] (last visited Feb. 15, 2021).

[89] A.B. A8155-B, 2017-2018 State Assemb., Reg. Sess. (N.Y. 2017). A parallel version of this bill originated and died in the New York State Senate around the same time. S.B. S5857, 2017-2018 S., Reg. Sess. (N.Y. 2018).

[90] *See, e.g.*, Eriq Gardner, *Disney Comes Out Against New York's Proposal to Curb Pornographic "Deepfakes*,*"* HOLLYWOOD REP. (June 11, 2018, 4:01 PM), https://www.hollywoodreporter.com/thr-esq/disney-new-yorks-proposal-curb-pornographic-deepfakes-1119170 [https://perma.cc/SSZ7-3HZS].

[91] S.B. S5959-D, 2019-2020 Leg., Reg. Sess. (N.Y. 2020) (enacted).

New York state legislature in the summer of 2020 and was subsequently signed by Governor Andrew Cuomo.[92]

Massachusetts. In 2019, legislation was introduced in the Massachusetts House of Representatives to "expand the state's definition of identity fraud to criminalize the creation or distribution of deepfakes intended for use in [or facilitation of] otherwise criminal or tortious conduct."[93] The bill defines a deepfake as "an audiovisual record," including photographs, videos, images, or sound recordings, that are "created or altered in a manner that the record would falsely appear to a reasonable observer to be an authentic record of the actual speech or conduct of an individual."[94] Rather than defining prohibited uses for deepfakes, such as nonconsensual pornography or election interference, the bill more broadly provides for liability if an individual creates or knowingly distributes a deepfake to commit or "facilitate criminal or tortious conduct."[95]

C. *Federal Legislation*

Thus far, federal bills have focused on generating studies and reports on deepfakes.[96] Such proposals recognize the importance of staying apprised of changes in deepfake technology. However, an effective law should anticipate technological changes, account for increasing sophistication, and create a framework that accommodates such changes. More importantly, a law should actually regulate malicious behavior that results in the harms discussed above.[97]

In 2018, Senator Ben Sasse introduced the Malicious Deep Fake Prohibition Act of 2018,[98] which would criminalize creating or knowingly distributing a deepfake with the intent to "facilitate criminal or tortious conduct under Federal, State, local, or Tribal law."[99] Like the bill proposed in Massachusetts, the Malicious Deep Fake Prohibition Act does not prohibit specific uses of

[92] N.Y. CIV. RIGHTS LAW § 50-F (McKinney 2021); N.Y. CIV. RIGHTS LAW § 52-C (McKinney 2021).

[93] FERRARO, *supra* note 73, § III.

[94] H.B. 3366, 191st Gen. Court, 1st Ann. Sess. § 1 (Mass. 2019).

[95] *Id.*

[96] FERRARO, *supra* note 73, § I. Various bills have been introduced that would require, among other things: (1) periodic reports by the Department of Homeland Security on changes in deepfake technology and laws governing deepfakes; (2) periodic briefings of congressional defense committees from the Secretary of Defense on deepfake technology, detection, and threats; (3) reports from the Director of National Intelligence on the national security impacts of deepfakes; and (4) a study from the Secretary of Defense on the cyberexploitation of military members. *Id.*

[97] *See supra* Part I (describing dangers that deepfakes pose to national security and election integrity).

[98] S. 3805, 115th Cong. (2018).

[99] *Id.* § 2.

deepfakes, instead creating additional grounds for liability for facilitating already illegal or tortious conduct.[100]

In 2019, Representative Clarke introduced the most expansive piece of federal deepfake legislation, the DEEP FAKES Accountability Act.[101] This Act proposes labeling requirements for persons who produce an "advanced technological false personation record with the intent to distribute such record over the internet."[102] A person who knowingly fails to disclose that audiovisual content is a deepfake or alters the mandatory disclosure requirements with the intent to distribute a deepfake faces criminal and civil liability, including up to five years imprisonment and a fine.[103]

Whoever violates this subsection faces liability if they acted: (1) "with the intent to humiliate or otherwise harass the person falsely exhibited" if the deepfake contains "sexual content"; (2) "with the intent to cause violence or physical harm, incite armed or diplomatic conflict, or interfere in an . . . election" if the deepfake posed a "credible threat of instigating or advancing such"; (3) "in the course of criminal conduct related to fraud, including securities fraud and wire fraud, false personation, or identity theft"; or (4) "by a foreign power, or an agent thereof, with the intent of influencing a domestic policy debate, interfering in a Federal, State, local, or territorial election," or engaging in other such unlawful acts.[104]

Importantly, the DEEP FAKES Accountability Act provides an exception for situations in which "a reasonable person would not mistake the falsified material activity for actual material activity of the exhibited living person," including parodies, historical reenactments, and fictionalized radio, television, or movies.[105]

III. CONSTITUTIONAL CONSTRAINTS ON REGULATING DEEPFAKES

Enacting legislation designed to prevent deepfakes from creating national security harms and interfering with elections would pose several constitutional issues. Regulation of deepfakes is regulation of speech that implicates foundational rights to self-expression under the First Amendment.[106] Moreover, legislation that attempts to regulate speech and expression using a national security rationale must satisfy an additional line of Supreme Court precedent

[100] *Id.*

[101] H.R. 3230, 116th Cong. (2019).

[102] *Id.* § 2. The DEEP FAKES Accountability Act would require the creator of a deepfake to provide an embedded digital watermark and audio and visual disclosures that the content contains altered audio and or visual elements. *Id.*

[103] *Id.*

[104] *Id.*

[105] *Id.* The DEEP FAKES Accountability Act also exempts from liability deepfakes "produced by an officer or employee of the United States, or under the authority thereof, in furtherance of public safety or national security." *Id.*

[106] *See* U.S. CONST. amend I.

that requires the Court to balance identifiable threats against other competing interests.[107] These constitutional limits must inform any attempt by Congress to regulate deepfakes that maliciously target elections or national security.

A. *First Amendment Doctrine*

A prohibition or selective ban on deepfakes, which are essentially a form of false speech, necessarily implicates First Amendment rights to free speech and expression.[108] Prior to 2012, the Supreme Court appeared to suggest that false statements lacked strong constitutional protections.[109] However, in 2012, a plurality of the Court concluded in *United States v. Alvarez*[110] that "falsity alone may not suffice to bring the speech outside the First Amendment" and that there has never been a "categorical rule . . . that false statements receive no First Amendment protection."[111] Under the First Amendment, the government can regulate false statements where one intends to cause "legally cognizable harm" and there is a causal link between "the restriction imposed and the injury to be prevented."[112] Thus, false statements are likely not protected by the First Amendment where they have a "propensity to bring about serious harms and [a] slight contribution to free speech values."[113] The *Alvarez* Court provided examples of such unprotected lies, including false statements to government officials, perjury, and impersonating government officials.[114]

[107] *See* Geoffrey R. Stone, *Free Speech and National Security*, 84 IND. L.J. 939, 950 (2009) (summarizing Cold War–era precedent).

[108] Cass R. Sunstein, *Falsehoods and the First Amendment*, 33 HARV. J.L. & TECH. 387, 421 (2020) [hereinafter Sunstein, *Falsehoods*] (discussing constitutional boundaries to regulating deepfakes and proposing that "[t]he government can regulate or ban deepfakes, consistent with the First Amendment, if (1) it is not reasonably obvious or explicitly and prominently disclosed that they are deepfakes, and (2) they would create serious personal embarrassment or reputational harm" (emphasis omitted)).

[109] *See* Hustler Mag., Inc. v. Falwell, 485 U.S. 46, 52 (1988) ("False statements of fact are particularly valueless; they interfere with the truth-seeking function of the marketplace of ideas, and they cause damage to an individual's reputation that cannot easily be repaired by counterspeech, however persuasive or effective."); Sunstein, *Falsehoods*, *supra* note 108, at 390-91 n.17 (collecting Supreme Court cases from 1964 to 2003 that "suggest[] that false statements lack constitutional protection"); Alan K. Chen & Justin Marceau, *High Value Lies, Ugly Truths, and the First Amendment*, 68 VAND. L. REV. 1435, 1443 (2015) ("[I]t was assumed that false factual statements are of no value to public discourse and thus fall entirely outside of the First Amendment's protections. The Court's rhetoric was unequivocal on this point.").

[110] 567 U.S. 709 (2012) (plurality opinion).

[111] *Id.* at 719. Prior cases instructed that such a statement "must be a knowing or reckless falsehood" to fall outside the scope of the First Amendment. *Id.* (citing, *inter alia*, N.Y. Times Co. v. Sullivan, 376 U.S. 254, 280 (1964)).

[112] *Id.* at 719, 725.

[113] Chesney & Citron, *supra* note 21, at 1791.

[114] *See Alvarez*, 567 U.S. at 719-20.

The inherent need to balance serious harms against contributions to free speech values warrants a sliding-scale approach to determine whether deepfakes should receive First Amendment protection. Such a sliding scale could balance factors indicating potential harm—such as intent, persuasiveness, and believability—against factors indicating positive speech values, such as facilitation of public debate or parody. It is settled that certain deepfakes have a propensity to bring about serious harms and that using deepfakes to deceive viewers and create harm fails to comport with established policy concerns underlying free speech protections. Deepfakes also have the capacity to contribute to free speech values—as evidenced by their beneficial capabilities in artwork, education, and commercial settings—and these need to be balanced against the myriad harms they could create.[115] But the harms associated with deepfakes that seek to undermine national security interests or interfere with elections outweigh their minimal contributions to speech. Accounting for such variability through a sliding-scale, case-by-case approach would therefore uphold traditional free speech values while protecting individuals and society from malicious deepfakes.

The First Amendment's philosophical underpinnings further illuminate the tension between the malicious use of deepfakes and the Amendment itself. First, promoting malicious and persuasive deepfakes—falsehoods calculated to bring about harm—does not aid in the search for truth in the marketplace of ideas, where "the best test of truth is the power of the thought to get itself accepted in the competition of the market."[116] Far from aiding in a search for truth, persuasive deepfakes are tantamount to "[f]alse statements of fact" and therefore "interfere with the truth-seeking function of the marketplace of ideas."[117] Second, and especially important in the context of election interference, deepfakes undermine self-governance by casting doubt on the electoral process.[118] Finally, the malicious use of deepfakes undercuts self-fulfillment and autonomy interests by depicting individuals without their consent.[119] These policy considerations bolster the conclusion that persuasive, believable, and malicious deepfakes warrant minimal First Amendment protection under a sliding-scale approach.

[115] *See supra* note 27.

[116] Abrams v. United States, 250 U.S. 616, 630 (1919) (Holmes, J., dissenting).

[117] Hustler Mag., Inc. v. Falwell, 485 U.S. 46, 52 (1988).

[118] *See* Alexander Meiklejohn, *The First Amendment Is an Absolute*, 1961 SUP. CT. REV. 245, 255-56.

[119] David A.J. Richards, *Free Speech and Obscenity Law: Toward a Moral Theory of the First Amendment*, 123 U. PA. L. REV. 45, 62 (1974) ("[T]he significance of free expression rests on the central human capacity to create and express symbolic systems, such as speech, writing, pictures, and music."). In this manner, deepfakes fail to "nurture[] and sustain[] the self-respect of the mature person," undermining the "value of free expression, [which] . . . rests on its deep relation to self-respect arising from autonomous self-determination." *Id.*

B. *National Security Rationale*

While the First Amendment principles above generally apply to government regulation of speech, the Supreme Court has sometimes employed a modified analysis when national security interests are at stake. Historically, the Supreme Court has allowed national security concerns to override free speech interests in two contexts: (1) wartime dissent and government criticism and (2) government secrecy.[120] Determining whether a national security rationale justifies prohibition or limitation of speech involves a balancing of competing interests unique to each situation.[121] Justice Frankfurter cautioned against a mechanistic approach in his concurrence in *Dennis v. United States*,[122] writing that "[t]he demands of free speech in a democratic society as well as the interest in national security are better served by candid and informed weighing of competing interests, within the confines of judicial process, than by announcing dogmas too inflexible for the non-Euclidian problems to be solved."[123] Although deepfakes do not easily fall under the two primary categories listed above, the balance of the relevant competing interests instructs that they may be regulated in order to promote national security. As discussed above, such determinations would be best made through a sliding-scale approach that balances potential harms, malicious intent, persuasiveness, and deceptiveness against positive speech values.[124]

After considering relevant national security interests, legislation must be narrowly tailored to target an identifiable threat. In *United States v. Robel*,[125] the Court held that the Subversive Activities Control Act of 1950, which prohibited members of registered Communist organizations from working in defense facilities, violated the First Amendment because the Act swept "indiscriminately across all types of association with Communist-action groups, without regard to the quality and degree of membership."[126] The Court noted the validity of

[120] Stone, *supra* note 107, at 939. Neither of these special circumstances would be implicated by regulating deepfakes. A deepfake that merely criticized the government in a nonmisleading way would amount to pure political speech protected by the First Amendment. And although forgery of confidential documents is one harm posed by deepfakes, the government likely has no interest in maintaining secret documents that are not actually secret.

[121] Thomas M. Franck & James J. Eisen, *Balancing National Security and Free Speech*, 14 N.Y.U. J. INT'L L. & POL. 339, 342-43 (1982) (evaluating competing interests in national security and free speech in context of *Snepp v. United States*, 444 U.S. 507 (1980)).

[122] 341 U.S. 494 (1951).

[123] *Id.* at 524-25; *see also* United States. v. Robel, 389 U.S. 258, 264 (1967) ("When Congress' exercise of one of its enumerated powers clashes with those individual liberties protected by the Bill of Rights, it is our 'delicate and difficult task' to determine whether the resulting restriction on freedom can be tolerated." (quoting Schneider v. State, 308 U.S. 147, 161 (1939)).

[124] *See supra* text accompanying notes 114-115.

[125] 389 U.S. 258 (1967).

[126] *Id.* at 260-62.

congressional concern over the danger of sabotage and espionage in the defense industry, leaving the door open for Congress to write narrowly tailored legislation targeting that perceived threat while nonetheless striking down the Act.[127] Accordingly, for the *Robel* Court, an identifiable harm could eventually reach a tipping point that warranted government regulation over an individual right[128]: "[W]hile the Constitution protects against invasions of individual rights, it is not a suicide pact."[129] Scholars are divided on whether there will be, or already has been, such a tipping point in the modern online speech landscape where a perceived danger is so great that it may warrant limiting free speech under the guise of national security or another justification.[130]

Online terrorist advocacy, where terrorist cells can rapidly exhort followers via social media to commit violent acts, is a strong candidate for such a tipping point.[131] In arguing against content-based restrictions on speech in response to such threats, scholars have stressed the numerous instances where the United States compromised "First Amendment freedoms in the face of perceived danger" only to "later recogniz[e] that we had overreacted, often with dire consequences for individual freedom and for our democracy."[132] Infamous examples of this include prosecutions under the Sedition Act of 1917, the Espionage Act of 1917, and other criminal syndicalism statutes.[133]

The "broad emergence of the internet and social media" has fundamentally changed the effect of speech and may warrant a change in doctrine or how existing doctrine is applied.[134] In particular, some scholars have pointed to

[127] The Court explained in detail,

> We are not unmindful of the congressional concern over the danger of sabotage and espionage in national defense industries, and nothing we hold today should be read to deny Congress the power under narrowly drawn legislation to keep from sensitive positions in defense facilities those who would use their positions to disrupt the Nation's production facilities. We have recognized that, while the Constitution protects against invasions of individual rights, it does not withdraw from the Government the power to safeguard its vital interests.

Id. at 266-67.

[128] David S. Han, *Terrorist Advocacy and Exceptional Circumstances*, 86 FORDHAM L. REV. 487, 493 (2017).

[129] Kennedy v. Mendoza-Martinez, 372 U.S. 144, 160 (1963).

[130] *See* Han, *supra* note 128, at 493-94 (surveying scholarly perspectives on proposals to limit terrorist advocacy and whether "tipping point has already been reached").

[131] *Id.* at 494 ("[A] time may well come when courts will have to grapple with circumstances sufficiently severe to justify content-based regulations of abstract terrorist advocacy . . . .").

[132] Geoffrey R. Stone, *ISIS, Fear, and the Freedom of Speech*, HUFFPOST (Dec. 22, 2016), https://www.huffpost.com/entry/isis-fear-and-the-freedom_b_8864050 [https://perma.cc/WF7U-2XJ7].

[133] *Id.* (citing examples of ill-advised attempts to "restrict our most fundamental freedoms in moment[s] of panic").

[134] Han, *supra* note 128, at 495 ("[E]xceptional circumstances might also reflect deeper

dangers arising from exceptional circumstances, such as online terrorist advocacy, that warrant a shift in the doctrine.[135] For instance, social media "can dramatically amplify the capacity of speech in one place to cause violence elsewhere at some uncertain time."[136] And "[i]t is the change in technology, more than the change in the nature of foreign threats, that has given rise to a historic and unprecedented danger."[137]

When confronted with exceptional circumstances such as terrorist advocacy, scholars suggest that courts should rethink the way that they apply traditional First Amendment doctrine.[138] In such cases, the threshold question for determining if an exceptional circumstance exists is whether there is "an outsized degree of actual or estimated harm" to national security, which "will depend heavily on the origin and nature of the circumstances in question."[139] Reformulation is best suited for when the exceptional circumstance is the "product of deeply rooted social or technological changes that fundamentally alter the basic balance between speech protection and the government's regulatory interests."[140] Finally, any regulation curtailing First Amendment rights in furtherance of security interests would need to satisfy traditional strict scrutiny analysis, particularly for content-based regulations targeted at a special class of deepfakes.[141]

Parallels exist between online terrorist advocacy and the malicious use of deepfakes to undermine national security and interfere in elections. Both threats take advantage of the rapid deployment and circulation offered by the Internet and social media, threatening immediate harms at "some uncertain time."[142]

social or technological changes that are more far-reaching and permanent in nature.").

[135] *See* Eric Posner, *ISIS Gives Us No Choice but to Consider Limits on Speech*, SLATE (Dec. 15, 2015, 5:37 PM), https://slate.com/news-and-politics/2015/12/isiss-online-radicalization-efforts-present-an-unprecedented-danger.html [https://perma.cc/948A-8JRD] ("Never before in our history have enemies outside the United States been able to propagate genuinely dangerous ideas on American territory in such an effective way . . . . The novelty of this threat calls for new thinking about limits on freedom of speech."); Cass R. Sunstein, Opinion, *Islamic State's Challenge to Free Speech*, BLOOMBERG (Nov. 23, 2015, 12:38 PM) [hereinafter Sunstein, *Islamic State*], https://www.bloomberg.com/opinion/articles/2015-11-23/islamic-state-s-challenge-to-free-speech (arguing that the "clear and present danger test," which forbids the government from regulating speech without clear and imminent danger, should be reconsidered in the age of terrorist advocacy and social media).

[136] Han, *supra* note 128, at 495 (quoting Sunstein, *Islamic State*, *supra* note 135).

[137] Posner, *supra* note 135.

[138] Han, *supra* note 128, at 494-97 (arguing that exceptional circumstances such as terrorist advocacy may "justify a radical departure from the robust constitutional protection broadly afforded to abstract advocacy").

[139] *Id.* at 497.

[140] *Id.* at 498.

[141] *Id.* (arguing strict scrutiny analysis "should *always* be the first-line test to account for these sorts of exceptional circumstances").

[142] Sunstein, *Islamic State*, *supra* note 135 (questioning applicability of the "clear and

Moreover, both threats are exacerbated as online communications technology becomes easily and widely accessible. The difference with deepfakes is that the resulting harm stems from their falsity and the context in which they are produced and distributed. National security precedent, paired with scholarly insights about the potential regulation of violent and extremist content, modernizes the framework with which to address free speech regulations that target perceived national security threats.[143]

As discussed above, prohibiting or limiting speech on the basis of national security involves weighing competing interests.[144] Such legislation likely must target an identifiable and probable threat through narrowly tailored means.[145] Accordingly, legislation regulating deepfakes targeting national security or election interference must account for both the harm to be avoided and the risk of impeding constitutionally protected speech. Because deepfakes have beneficial uses, an outright ban would be overbroad.[146] In considering an identifiable and probable threat, a sliding-scale approach should consider such factors as the intent, persuasiveness, and believability of a deepfake. Narrowly targeted legislation based on a sufficient showing of harm would satisfy precedent governing falsehoods under the First Amendment.[147]

Thus, to warrant prohibition, deepfakes must be believable and designed to cause harm. Prohibiting demonstrably fake videos that do not actually persuade or deceive viewers or those that are made to parody or satire would chill free speech and violate existing precedent. However, a highly persuasive deepfake that actually deceives viewers and is intended to influence a voter's decision or undermine national security would likely fall outside the bounds of protected speech. Moving forward, the specific harms posed by widespread circulation of increasingly sophisticated deepfakes warrant a careful application of First Amendment precedent, an area in which courts have been traditionally hesitant to limit free speech and expression.

present danger test" where social media could incite violence at "some uncertain time").

[143] *See, e.g.*, Han, *supra* note 128, at 494-97; Posner, *supra* note 135; Sunstein, *Islamic State*, *supra* note 135.

[144] *See* Franck & Eisen, *supra* note 121, at 343 ("A diligent court would ask whether there are weightier countervailing interests.").

[145] Han, *supra* note 128, at 497-98.

[146] Overbreadth is a common justification for finding regulations of free speech unconstitutional. *See, e.g.*, Virginia v. Black, 538 U.S. 343, 365-67 (2003) (plurality opinion) (invalidating portion of a cross-burning statute that regarded all cross burning as prima facie evidence of an intent to intimidate because it unconstitutionally prohibited too much speech, specifically that not intended to intimidate).

[147] Sunstein, *Falsehoods*, *supra* note 108, at 421-22 ("Under *Alvarez*, there should be no constitutional barrier to allowing controls on deepfakes, at least on a sufficient showing of harm . . . . Those controls might take the form of a regulatory approach, operating perhaps via an independent commission, or (more interestingly) a tort-like approach, operating through a civil cause of action, building on libel law, and creating a kind of property right in one's person.").

IV. NARROWLY TAILORED LEGISLATION

A targeted solution that combats threats to national security and election integrity is necessary to deter malicious actors who, at the moment, "face no real consequences for creating videos [and other content] that [is] hugely destructive to our societies."[148] The current state of the law, which consists of scattered criminal and civil liability at the state level, is insufficient to stop the threatened harms.[149] A coherent solution requires consistency and a flexible, forward-looking approach that can anticipate and adapt to new iterations of deepfake technology. Identifying a silver bullet that stops malicious deepfakes in their path is likely impossible. However, existing and proposed state and federal laws provide a starting point from which to design a comprehensive and effective federal solution.[150] Desirable aspects from existing and proposed laws, which should form the basis for such a solution, can be divided into three categories: (1) prohibited acts, (2) primary and secondary liability, and (3) mechanisms for bringing lawsuits or removing content.

A. *Prohibited Acts and Associated Harms*

At the outset, legislation that targets malicious deepfakes that threaten national security and election integrity must comport with the First Amendment. Indeed, it must be narrowly tailored legislation that targets such harms and abides by free speech and national security precedent to be constitutionally permissible.[151] Moreover, in regulating national security concerns, legislation must properly balance competing government interests with free speech interests.[152] Thus, limits on free speech, selective prohibitions, or regulations of deepfakes are only justified in specific circumstances: (1) if they are narrowly tailored to address a legally cognizable injury and (2) if there is a causal connection between the limitation and the harm.[153] Under this framework, any limitation must be causally linked to the specific injuries flowing from deepfakes: (1) the speaker's intent to cause harm due to the false nature of the speech and (2) the persuasiveness and believability of the speech within the context in which it is spoken.[154] This balancing, or sliding-scale, approach

[148] Clarke, *supra* note 29.

[149] *Id.* ("Deepfakes are a threat to the truth on which we base our democracy.").

[150] As it proposed to do in the DEEP FAKES Accountability Act, Congress could regulate deepfakes under the Commerce Clause. *See* H.R. 3230, 116th Cong. § 2 (2019).

[151] Chesney & Citron, *supra* note 21, at 1791-92.

[152] *See* Franck & Eisen, *supra* note 121, at 342-43 ("Essentially, a determination must be made as to which costs may be rightfully imposed in light of more important benefits.").

[153] United States v. Alvarez, 567 U.S. 709, 725 (2012) (plurality opinion) (holding that the Stolen Valor Act infringed on respondent's First Amendment rights where government failed to show "direct causal link" between interest in upholding integrity of military honors system and Act's prohibition against false claims of receiving Medal of Honor).

[154] *See supra* Part III (discussing when and how false statements may be regulated consistently with First Amendment).

adheres to established doctrine and ensures flexibility in applying legislation and adjudicating outcomes. Indeed, existing and proposed state and federal legislation appears to have been crafted with some of these interests in mind.

1. Intent to Cause Harm

Out of the patchwork of existing and emerging laws, nearly every scenario requires that a creator or distributor of a deepfake has the specific intent to bring about a specific harm.[155] With respect to national security threats and election interference, the DEEP FAKES Accountability Act targets specific harms by mandating an "intent to distribute" the deepfake in addition to an "intent to cause violence or physical harm, incite armed or diplomatic conflict, or interfere in an official proceeding, including an election."[156] Importantly, this Act requires a causal link between the speech and the resulting injury, such that "[the deepfake] did in fact pose a credible threat of instigating or advancing such [harms]."[157] Further addressing election interference, the Act establishes liability where "a foreign power" intends to influence domestic policy or interfere "in a Federal, State, local, or territorial election."[158] By requiring that the deepfake poses a credible threat to specified harms, this solution is narrowly tailored enough so as not to offend the First Amendment.[159] Moreover, examining credibility through the context and circumstances in which each individual deepfake is produced or distributed ensures further flexibility by establishing a case-by-case approach. Thus, existing legislation provides a strong starting point for targeting cognizable injuries and acts in the national security and election integrity contexts.

Existing legislation could be modified by broadening its scope to include acts not specifically enumerated but which would be prohibited because of their likelihood to bring about similar harms and injuries. Congress could expand liability by prohibiting the creation or distribution of deepfakes to "facilitate criminal or tortious conduct," as has been proposed in Massachusetts.[160] On the one hand, such language broadens the scope of the legislation without needing to anticipate and enumerate specific harms, effectively remedying the uncertainty created by producing or distributing deepfakes. However, such catchall provisions risk making this legislation overbroad and at odds with First Amendment and national security precedent. Nonetheless, as long as any all-encompassing language is sufficiently narrowed by requiring a link between

[155] *See supra* Part II.

[156] H.R. 3230, 116th Cong. § 2 (2019).

[157] *Id.*

[158] *Id.*

[159] *See* Chesney & Citron, *supra* note 21, at 1791.

[160] H.B. 3366, 191st Gen. Court, 1st Ann. Sess. § 1 (Mass. 2019) ("Whoever . . . creates, with the intent to distribute, a deep fake and with the intent that the distribution of the deep fake would facilitate criminal or tortious conduct . . . shall be guilty of the crime of identity fraud . . . .").

intent and specified harms to national security and election integrity,[161] the broadened scope could fill potentially unanticipated gaps in legislation. In sum, an effective solution must, at a minimum, identify specified harms but could be improved by a catchall provision that allows for flexibility in a rapidly evolving technological landscape.

2. Persuasiveness and Believability

Simply put, deepfakes targeted at undermining democratic institutions and creating violence are damaging to the extent that they are persuasive and believable.[162] Seemingly authentic videos, images, and audio can advance false and damaging narratives or confirm preexisting cognitive biases about contested issues, which can then be harnessed to advance conspiracy theories.[163] In the national security context, unrealistic and implausible deepfakes would not have the same capacity as hyperrealistic deepfakes to disseminate false information, exploit social division, or create political unrest.[164] Similarly, deepfakes are most likely to interfere with an election when they are released close to the election itself, creating a "decisional chokepoint[]" and not giving candidates or fact-checkers enough time to rebut the false media.[165]

Accordingly, in addressing specific harms caused by malicious deepfakes, an effective legislative solution must, perhaps above all else, focus on the deepfakes' persuasiveness and believability. Indeed, several proposed and enacted solutions have already taken these concerns into account. For instance, as a threshold matter, pieces of proposed legislation often adopt a reasonable person standard, which asks whether a reasonable person would believe that the manipulated audiovisual material was in fact authentic.[166] Moreover, the DEEP FAKES Accountability Act includes mandatory watermarking and disclosure

[161] *See* H.R. 3230 § 2 (requiring direct causal connection between deepfake and potential harm to trigger penalties).

[162] *See* Nina I. Brown, *Deepfakes and the Weaponization of Disinformation*, 23 VA. J.L. & TECH. 1, 9-13 (2020) (describing how deepfakes create "war on reality" by confirming preconceived notions and causing individuals to question authenticity of any audiovisual record). In other contexts, such as nonconsensual pornography, deepfakes are damaging both because they are perceived as authentic and because they cause humiliation, trauma, and reputational harms. *See* Chesney & Citron, *supra* note 21, at 1773-74.

[163] *See* Chesney & Citron, *supra* note 21, at 1777-78, 1785-86; *see also* Brown, *supra* note 162, at 10-11.

[164] *See* Schiff, Murphy & Curbelo, *supra* note 45.

[165] *See* Chesney & Citron, *supra* note 21, at 1778-79.

[166] *See, e.g.*, H.R. 3230 § 2 (granting exemption from liability where "reasonable person would not mistake" the deepfake's falsified content for legitimate content, such as "parody shows" or "historical reenactments"); *see also* CAL. ELEC. CODE § 20010(e) (West 2020) (defining "materially deceptive audio or visual media" as that which "falsely appear[s] to a reasonable person to be authentic" and causes "a reasonable person to have a fundamentally different understanding or impression" of the deepfake's content than had person seen the "unaltered, original version").

provisions to guard against the risk that viewers would think that a deepfake was real.[167] In the context of election interference, several proposals address deepfakes' persuasiveness by holding individuals liable only within a certain time period before an election[168]—this temporal limit allows for greater creative expression through deepfake technology but narrowly targets instances when the deepfakes could do the most harm. In creating a comprehensive solution, effective legislation should therefore guard against the deleterious effects of persuasiveness and believability by mandating disclosures that dispel confusion, adopting a reasonable person standard to determine whether a viewer is deceived, and allowing for context-specific considerations, such as timing before an election.

Several of these concepts, such as requiring a disclosure to dispel the confusion of whether media is a deepfake, have already been incorporated in proposed legislation such as the DEEP FAKES Accountability Act. The DEEP FAKES Accountability Act requires watermarks and other disclosures on "[a]ny advanced technological false personation record which contains a moving visual element."[169] Importantly the Act anticipates digital, "audiovisual," "visual," and "audio" deepfakes, requiring either a digital watermark identifying the media as modified, or clear "written" or "verbal statement[s]" identifying the degree of the media's modification.[170] Under this disclosure requirement, wrongdoers are liable for either omitting or altering the necessary disclosure.[171] Such disclosures are vital to addressing the unique harms posed by persuasive and believable deepfakes. If properly used, mandated watermarks and disclosures would foster trust and reliability by giving viewers some indication of whether a picture, video, or audio recording is real or fake.[172]

However, such disclosure requirements could also play into a wrongdoer's hands. After all, a malicious actor is unlikely to comply with disclosure requirements,[173] and a viewer primed to believe the authenticity of any media not containing a watermark or disclosure would therefore be easily duped. On

[167] H.R. 3230 § 2.

[168] For example, Texas has a thirty-day window before elections while California's proposed legislation has a sixty-day window. *See* TEX. ELEC. CODE ANN. § 255.004(d)(2) (West 2019); CAL. ELEC. CODE § 20010(a) (West 2020).

[169] H.R. 3230 § 2.

[170] *Id.*

[171] *Id.*

[172] *See* Chesney & Citron, *supra* note 21, at 1786 (explaining the "combination of *truth* decay and *trust* decay" that arises when deepfakes frustrate public's belief in "what their eyes or ears are telling them—even when the information is real").

[173] Similar nihilistic arguments are often advanced in the gun control debate: Criminals do not follow laws, so why have gun laws at all? *See, e.g.*, *Study Reinforces What We Already Know: Criminals Don't Follow the Law*, NRA INST. FOR LEGIS. ACTION (June 28, 2019), https://www.nraila.org/articles/20190628/study-reinforces-what-we-already-know-criminals-don-t-follow-the-law [https://perma.cc/6YE2-L8JF].

balance, however, disclosures would likely lead to increased trust and stability by allowing the viewer to distinguish between real and fake content[174] and conditioning viewers to seek out such disclosures. Particularly because many harms associated with deepfakes stem from mistrust and manipulation, a legal solution should strive to inform viewers about the authenticity of a picture, video, or audio recording.

Another tactic for addressing deepfakes' persuasiveness and believability is to adopt a reasonable person standard for what constitutes a deepfake and whether someone would be deceived. The DEEP FAKES Accountability Act adopts a "reasonable person" standard in defining an "advanced technological false personation record" as "any deep fake, which . . . a reasonable person, having considered the visual or audio qualities of the record . . . would believe accurately exhibits" prohibited content.[175] In addition, California's Elections: Deceptive Audio or Visual Media Act applies to images, videos, and audio recordings that "falsely appear to a reasonable person to be authentic" and "cause a reasonable person to have a fundamentally different understanding or impression of the expressive content . . . than that person would have if the person were hearing or seeing the unaltered, original version."[176] Under both proposals, no one is liable for creating or distributing an unrealistic deepfake. Deepfakes designed to create violence or interfere with elections are damaging to the extent that they are believable and persuasive. California's proposal is therefore particularly strong because it specifically addresses both of these considerations. If a deepfake does not "falsely appear to a reasonable person to be authentic," it is not believable; a deepfake that does not "cause a reasonable person to have a fundamentally different understanding" of the altered content is not persuasive.[177] A solution incorporating the DEEP FAKES Accountability Act and California's Elections: Deceptive Audio or Visual Media Act is therefore desirable to combat those deepfakes most likely to result in tangible electoral or national security harms.

The reasonable person standard proposed in the DEEP FAKES Accountability Act and California's Elections: Deceptive Audio or Visual Media Act also addresses First Amendment concerns by creating exceptions for deepfakes that constitute satire and parody.[178] A satire-and-parody exception is important from a doctrinal perspective because it reduces the law's scope; satire

[174] *See* Brown, *supra* note 162, at 11.

[175] H.R. 3230 § 2.

[176] CAL. ELEC. CODE § 20010(e) (West 2020).

[177] *Id.* § 20010(e).

[178] H.R. 3230 § 2 (asserting that deepfakes developed as "parody shows or publications, historical reenactments, or fictionalized radio, television, or motion picture programming" do not require disclosure when "a reasonable person would not mistake" false content for truthful content); CAL. ELEC. CODE § 20010(d)(5) (West 2020) ("This section does not apply to materially deceptive audio or visual media that constitutes satire or parody.").

and parody enjoy extensive First Amendment protection.[179] Under the framework discussed above, the exception is desirable, and perhaps constitutionally necessary, because it allows for unpersuasive and unrealistic deepfakes. If a reasonable person realizes that a deepfake is satirical or parodical, they are neither persuaded nor deceived by its content—thus, the attendant harm from a malicious deepfake would not ensue. As already anticipated by the DEEP FAKES Accountability Act and California's Elections: Deceptive Audio or Visual Media Act, an effective solution must therefore incorporate an exception for parody and satire.

Finally, an effective solution must allow for flexibility by considering the context in which a particular deepfake is created and disseminated. Such an approach is practicable from both a common-sense perspective and doctrinal perspective. Indeed, the First Amendment and national security rationales demand balancing harm against positive contributions to speech.[180] Both the Texas Deepfake Act and California's Elections: Deceptive Audio or Visual Media Act accordingly hold wrongdoers liable for election interference only within a certain amount of time before an election.[181] A mandated window of liability inherently considers the context in which a particular deepfake is created and disseminated. As noted above, deepfakes have the greatest potential to interfere with an election when they are released without enough time to debunk or refute their content.[182] A deepfake released thirty or even sixty days before an election is less likely to interfere with the outcome because it can be fact-checked and refuted with counterspeech. Therefore, an indefinite ban in the election interference context may be overbroad and chill more speech than necessary.[183] An effective solution should undertake a similar analysis for each

[179] *See, e.g.*, Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579 (1994) ("Like less ostensibly humorous forms of criticism, [parody] can provide social benefit, by shedding light on an earlier work, and, in the process, creating a new one. We thus line up with the courts that have held that parody, like other comment or criticism, may claim fair use under [the Copyright Act]."); Hustler Mag., Inc. v. Falwell, 485 U.S. 46, 57 (1988) ("[T]his [libel] claim cannot, consistently with the First Amendment, form a basis for the award of damages when the conduct in question is the publication of a caricature such as the ad parody involved here.").

[180] *See* Chesney & Citron, *supra* note 21, at 1785-86 (discussing weighing harm against contribution to speech); Franck & Eisen, *supra* note 121, at 343 (describing balancing government interests against harm in national security context); Sunstein, *Falsehoods*, *supra* note 108, at 421-22 (arguing that deepfakes can be regulated upon sufficient showing of harm).

[181] *See* TEX. ELEC. CODE ANN. § 255.004(d)(2) (West 2019); CAL. ELEC. CODE § 20010(a) (West 2020).

[182] *See supra* text accompanying notes 63-64, 165; *see also* Chesney & Citron, *supra* note 21, at 1776-78 (discussing deepfake sabotage and harm it poses to society).

[183] Of course, elections are influenced by events that take place long before the election date. This analysis is meant to demonstrate how a legislative solution must consider the

facet of the legislation by carefully considering the context in which a deepfake is created and distributed, weighing harm against contributions to speech, and assessing the deepfake's persuasiveness and believability.

A seemingly unrelated but strikingly similar area of law offers a helpful template in crafting substantive, narrowly tailored legislation: trademark infringement through brandjacking on social networks.[184] Like deepfakes, brandjacking involves the impersonation of a likeness online, implicating First Amendment protections of freedom of expression.[185] Moreover, brandjacking involves substantial harms to individuals, such as harm to reputation. This arises in the trademark setting when an alleged infringer posts untruthful, offensive, or inappropriate information, creating confusion over the source of the expression.[186]

A permissible solution that balances trademark and free speech interests is to prohibit unauthorized use of a mark on social media where it is "likely to cause confusion about the source of expression unrelated to the advertising or sale of goods or services" and where (1) the mark impersonates the markholder and "falsely suggest[s] the markholder is the author of the third party's expression," (2) a reasonable person would believe the brandjacker's "false statements of identity and authorship," and (3) the content of the social media site "does not dispel the confusion regarding the source of the expression."[187] Like harms from malicious deepfakes, such an approach against brandjacking would mitigate persuasiveness and believability by adopting a reasonable person standard, mandating disclosure, and addressing the specific context in which brandjacking would confuse a consumer. In continuing to craft effective legislation, lawmakers should consider unrelated yet similar ideas like brandjacking. Creative solutions are needed for such a novel problem.

B. *Holding Creators and Distributors Liable*

The next step, after fashioning the substance of the legislation, is considering its remedial scope. This includes determining which individuals or entities are primarily or secondarily liable for the production and dissemination of

context in which deepfakes are produced and distributed to determine their persuasiveness and believability.

[184] Brandjacking is the illegal use of trademarked brand names online, such as Facebook and Twitter. *See* Lisa P. Ramsey, *Brandjacking on Social Networks: Trademark Infringement by Impersonation of Markholders*, 58 BUFF. L. REV. 851, 855 (2010) (discussing confusion created by brandjacking, and noting that perpetrators are likely to hide behind First Amendment free expression ideals).

[185] *Id.*

[186] *Id.* at 856 ("If the accused infringer is using the mark to impersonate the markholder and cause confusion about the source of expression on the social network site, some courts may find infringement even where the third party is not advertising or selling goods or services.").

[187] *Id.* at 859.

prohibited deepfakes and which individuals or entities can bring suit. Determining the scope of primary and secondary liability should be guided by the extent to which deepfakes aimed at disrupting elections and threatening violence create harm. In each context, although the harm exists once the deepfake is posted,[188] that harm only reaches its full potential when the deepfake is broadly circulated.[189] Such an approach fits within the framework discussed above, weighing the intent with which a deepfake is created and spread, its persuasiveness, and its believability against its positive contributions to speech. For instance, a deepfake created and privately kept by an individual without intent to cause harm or intent to distribute its content online lacks any level of persuasiveness because people cannot be persuaded by something they have never heard or seen.

1. The Problem with Primary Liability

It is clear that malicious deepfake producers should be held liable as primary actors—but for their production, resulting harms would not occur. In its current state, the DEEP FAKES Accountability Act holds liable persons who "produce[] an advanced technological false personation record with the intent to distribute such record over the internet or knowledge that such record shall be so distributed."[190] Thus, individual creators could be held liable for producing a prohibited deepfake with the intent to distribute it via the Internet, provided that other conditions are met. However, it should be noted that identifying and locating individual content creators on the Internet can be exceedingly difficult because these individuals can use sophisticated technologies to remain anonymous.[191] Deepfake creators may also avoid liability if they reside outside of the United States.[192] Because these content creators can escape liability by remaining anonymous, they are also difficult to deter and may be "effectively

[188] *See* Brown, *supra* note 162, at 14-15.

[189] *Id.* at 14-15 ("Deepfakes designed to disrupt elections or threaten public safety, for example, would necessarily rely on wide distribution in order to have their desired impact." (footnote omitted)).

[190] H.R. 3230, 116th Cong. § 2 (2019).

[191] *See* DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE 142-43 (2014) (explaining how Internet content creators use technological methods to remain anonymous and avoid detection). Creators can remain anonymous by using technologies like Tor or hiding their IP address. Chesney & Citron, *supra* note 21, at 1792; *see also* Rae Hodge, *Tor Browser FAQ: What Is It and How Does It Protect Your Privacy?*, CNET (Feb. 1, 2021, 6:00 AM), https://www.cnet.com/how-to/what-is-tor-your-guide-to-using-the-private-browser/ [https://perma.cc/BE2J-2S4K] (describing mechanics of Tor, which makes user's online activity nearly untraceable by relaying user's internet traffic many "relay nodes" before releasing it to the open Internet).

[192] Chesney & Citron, *supra* note 21, at 1792.

judgment-proof"; serving process and initiating a lawsuit may also be prohibitively expensive for plaintiffs.[193]

In the short term, online platforms themselves are best situated to deter.[194] Websites and social media platforms have already responded to disinformation spread through fake news and fake accounts by "building algorithms to 'contextualize' news with other sources" and removing fake accounts, among other efforts.[195] A legislative solution could endeavor to deter individual content creators in this manner. The DEEP FAKES Accountability Act, for example, contains a provision addressing private sector collaboration, which pledges that if the U.S. government "develops technology to reliably detect deep fakes," it will share that technology with online platforms.[196] This provision is too vague to be an effective solution, as legislation should collaborate with online platforms in holding individual content creators accountable. To the extent possible, a collaborative solution may require online platforms to moderate prohibited content or face secondary liability.

2. Secondary Liability

Although identifying individual creators will likely be difficult, that is not to say that a legislative solution must necessarily lack teeth. Deepfakes designed to undermine national security and interfere with elections rely on widespread dissemination often available through online platforms. Expanding secondary liability to online publishers or platforms may therefore be necessary given how deepfakes metastasize via the Internet—in particular, social media—and how difficult their creators can be to locate.[197] Moreover, such secondary liability is necessary because even if an individual is held liable, private companies may refuse to remove deepfakes from their platforms.[198] Of course, if a malicious

[193] *Id.* at 1792-93.

[194] *Id.* at 1795 ("In some contexts, [imposing liability on platforms] may be the only realistic possibility for deterrence and redress."); *see also* Brown, *supra* note 162, at 57 (explaining that social platforms "have the financial resources and technological expertise to contribute and much to lose if user trust in their platforms continues to erode").

[195] Brown, *supra* note 162, at 57 (quoting Justin Sherman, *Fighting Deepfakes Will Require More Than Technology*, NEXTGOV (Dec. 14, 2018), https://www.nextgov.com/ideas/2018/12/fighting-deepfakes-will-require-more-technology/153530/ [https://perma.cc/CQ7P-746A]).

[196] H.R. 3230, 116th Cong. § 7 (2019).

[197] Chesney & Citron, *supra* note 21, at 1762 (describing how social media accelerates spread of deepfakes).

[198] *E.g.*, David Gilbert, *Facebook Refuses to Remove Deepfakes of Zuckerberg, Trump and Kardashian*, VICE NEWS (June 12, 2019, 8:05 AM), https://www.vice.com/en_us/article/9kxgj3/facebook-refuses-to-remove-deepfakes-of-zuckerberg-trump-and-kardashian [https://perma.cc/7VLP-SH68]. Facebook has since changed its position and now says "it will ban videos that are heavily manipulated by artificial intelligence." David McCabe & Davey Alba, *Facebook Will Ban 'Deepfakes*,*'* N.Y. TIMES, Jan. 8, 2020, at B7.

deepfake is kept online, resulting injuries would ensue even if its creator is held liable. The same harm could occur if the original producer and distributor of a prohibited deepfake is unknown or if the altered video appears on a website or platform without a clear publisher. Accordingly, to mitigate the harm posed by deepfakes, legislatures should prioritize both deterring malicious actors and removing online content.

To create secondary liability, legislation should create rights of action to bring lawsuits against both creators and distributors for damages, as exemplified by California's efforts to target deepfakes.[199] In particular, California's Elections: Deceptive Audio or Visual Media Act would hold any entity liable that distributes prohibited content to interfere with an election.[200] Because recovery is often limited due to the difficulty of finding creators, victims[201] of systemic harms would likely be better situated if they are able to sue distributors of deepfakes, including online platforms.[202] In this way, online platforms are more likely to be deterred as the lowest-cost avoiders and, unlike most deepfake creators, they are not judgment proof.[203] Moreover, Internet service providers and online platforms have economic, moral, and market incentives to moderate and remove prohibited content.[204] Thus, holding distributors secondarily liable could effectively mitigate the spread of malicious deepfakes via online platforms and social media, which would directly address harms from deepfakes designed to create conflict and interfere with elections.

Broad societal harm stemming from a malicious, well-timed deepfake is only possible so long as the offending deepfake remains online. The greater the reach of a deepfake, the more likely it is to disrupt an election or to create violence and conflict.[205] In this context, an online platform or publisher could be equally culpable as content creators for resulting harm if the platform learns that it is hosting a malicious deepfake and refuses to take steps to remove the content.[206]

[199] CAL. ELEC. CODE § 20010 (West 2020) (criminalizing distribution of such deepfakes); Assemb. B. 1280, 2019-2020 Leg., Reg. Sess. (Cal. 2019) (criminalizing production of such deepfakes).

[200] CAL. ELEC. CODE § 20010 (West 2020).

[201] Note, however, that it is difficult to identify a particular "victim" to bring suit where the harm is a threat to national security. *See infra* Section IV.C. In contrast, a political candidate victimized by a malicious deepfake in an election is an easily identified and cognizable victim to bring suit, though attendant harms also affect society as a whole. *See infra* text accompanying notes 222-223.

[202] Chesney & Citron, *supra* note 21, at 1795.

[203] *See id*. ("In some contexts, this may be the only realistic possibility for deterrence and redress.").

[204] Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1616-17 (2018).

[205] *See* Brown, *supra* note 162, at 14-15 (discussing importance of wide distribution to effective use of malicious deepfakes).

[206] Defining and grappling with a knowledge standard for online platforms and publishers

Thus, to properly address the harms posed by deepfakes, a solution should strive to incorporate a provision like California's Elections: Deceptive Audio or Visual Media Act that holds distributors liable.

However, Section 230 of the Communications Decency Act largely immunizes online platforms from lawsuits for hosting harmful content that they did not themselves create.[207] Section 230 would thus likely need to be amended in order to hold distributors, including online platforms, accountable.[208] One suggestion is to amend the scope of Section 230 to hold liable online platforms that fail to take "reasonable steps to prevent or address unlawful uses of its services . . . as the publisher or speaker."[209] Such an amendment would mean that entities who fail to remove specifically prohibited unlabeled deepfakes, after receiving notice of them, could be held liable. A common objection to the notice-and-takedown regime imposed on online platforms under Section 230 is the practical and financial burden of flagging, moderating, and removing overwhelming amounts of content.[210] However, content moderation programs are less costly when they adopt criteria implementing narrowly tailored, precise rules as opposed to vague standards.[211] A program for taking down prohibited deepfakes could undoubtedly reduce costs by adopting precise rules. Indeed, it would be a simple rule to make any media identified as a deepfake and not containing a mandatory disclosure or watermark subject to removal.

Amending Section 230 in the face of modern technological problems is not unprecedented. Section 230 was amended in 2018 by the Allow States and Victims to Fight Online Sex Trafficking Act ("FOSTA"), which "clarif[ied] that section 230 . . . does not prohibit the enforcement against providers and users of

is beyond the scope of this Note. A potential solution could involve a notice-and-takedown regime like the Digital Millennium Copyright Act, which requires an online provider to have both objective and subjective knowledge that they are hosting prohibited content. 17 U.S.C. § 512 (laying out limitations in liability relating to material online). Unlike existing notice-and-takedown regimes, it is unclear which individuals or entities would have the incentive or ability to request that content be removed.

[207] Chesney & Citron, *supra* note 21, at 1795-96.

[208] Amending or restricting protections for online platforms is beyond the scope of this Note, although it has been the subject of recent debate. *See, e.g.*, Gilad Edelman, *Republicans Make an Unlikely Closing Pitch: Amend Section 230*, WIRED (Oct. 27, 2020, 4:54 PM), https://www.wired.com/story/senate-section-230-hearing-zuckerberg-dorsey-pichai/; Derek Khanna, *The Law that Gave Us the Modern Internet—and the Campaign to Kill It*, ATLANTIC (Sept. 12, 2013), https://www.theatlantic.com/business/archive/2013/09/the-law-that-gave-us-the-modern-internet-and-the-campaign-to-kill-it/279588/.

[209] Danielle Keats Citron & Benjamin Wittes, *The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity*, 86 FORDHAM L. REV. 401, 419 (2017) (emphasis omitted).

[210] *See* Brown, *supra* note 162, at 28 (describing the costly implementation process for online platforms to sift through and identify prohibited content); Klonick, *supra* note 204, at 1632 (describing implementation difficulties of YouTube and Facebook's content moderation programs).

[211] *See* Klonick, *supra* note 204, at 1633.

interactive computer services of Federal and State criminal and civil law relating to sexual exploitation of children or sex trafficking."[212] Thus, carve outs to Section 230 are conceivable, particularly when they address online harms unforeseen by Congress when it enacted Section 230 in 1996.[213] Like all aspects of the legislative solution discussed in this Note, an amendment to Section 230 for deepfakes requires balancing competing interests.[214] An approach that weighs harms flowing from a deepfake's intent, persuasiveness, and believability against contributions to speech is in accordance with Section 230's role in facilitating free speech.[215] Ultimately, secondary liability for distributors of prohibited deepfakes is desirable because it directly addresses the unique harms posed by deepfakes designed to undermine public safety and election integrity.

C. *Bringing a Lawsuit or Removing Content*

The final component of a comprehensive legislative solution is authorizing individuals or groups to bring lawsuits against individual content creators or against publishers and distributors. Such considerations necessarily draw on the discussion of primary and secondary liability presented above. Determining when and how an individual or group can bring suit or remove content from a platform must continue to be guided by how deepfakes designed to interfere with elections and create violence produce harm. To effectively prevent harm, malicious deepfakes must be quickly detected and removed before they wreak havoc by reaching a broader audience.[216] However, an overly broad ability to bring suit, especially where an individual or entity does not suffer any clear injury,[217] risks chilling speech and expression.

A private right of action is clearly appropriate where a deepfake is designed to injure an individual.[218] Indeed, the DEEP FAKES Accountability Act provides a private right of action for injunctive relief for "[a]ny living individual or affiliated corporate or other entity who has been exhibited as engaging in

[212] Allow States and Victims to Fight Online Sex Trafficking Act of 2017, Pub. L. No. 115-164, 132 Stat. 1253.

[213] Brown, *supra* note 162, at 45 ("Despite its impact on the development of the Web 2.0, it is conceivable that Section 230 could be revised to create legal exposure for platforms on which deepfakes are spread.").

[214] *See supra* Section IV.A (noting that deepfake prohibitions must be narrowly tailored and appropriately balance government interests against free speech concerns).

[215] *See* Klonick, *supra* note 204, at 1604-05 (detailing history and development of Section 230).

[216] *See* Brown, *supra* note 162, at 26-27.

[217] *See* United States v. Alvarez, 567 U.S. 709, 725 (2012) (plurality opinion) ("There must be a direct causal link between the restriction imposed and the injury to be prevented.").

[218] *See* Brown, *supra* note 162, at 14 ("The mere existence of a video that depicts [an individual] engaging in acts they never engaged in, without their consent, is harmful, even when it is not distributed.").

falsified material activity in an advanced technological false personation record" against someone who fails to make adequate disclosure or violates the Act's labeling requirement.[219] In the election interference context, California's Elections: Deceptive Audio or Visual Media Act provides a private right of action to "[a] candidate for elective office whose voice or likeness appears in a" prohibited deepfake.[220] As opposed to the DEEP FAKES Accountability Act, California's statute creates a narrower option, precisely identifying the individual who is authorized to bring suit. With the goal of identifying specific harms to public safety and election integrity in mind, California's approach is desirable because it creates a right of action in one of those specific circumstances. The DEEP FAKES Accountability Act's provision risks being overly broad and thus chilling too much desirable protected speech.[221] Although maintaining flexibility is important, an effective solution must be specific and not open the floodgates for litigants.

Deepfakes intended to threaten public safety or interfere in elections necessarily involve broader societal harm that does not affect just a single person; therefore, an individual's right to bring a private action is less clear.[222] Although candidates for election, world leaders, military personnel, or others specifically targeted by a malicious deepfake would likely have the ability to bring a private right of action (or at least presumably meet the Article III standing prerequisites to do so),[223] the attendant harm for a deepfake is far greater for society than the harm inflicted on a single individual. Moreover, federal and state governments have greater incentives and abilities to bring an action to enjoin deepfakes created to cause societal harm than do individuals who face financial, reputational, and other barriers to litigation. Ultimately, to mitigate societal harm where an individual's right to bring suit is not clear, the only recourse may be to remove the online content.

Because deepfakes intended to undermine national security and election integrity "necessarily rely on wide distribution in order to have their desired impact," there must be an efficient process for individuals or entities to remove

[219] H.R. 3230, 116th Cong. § 2 (2019).

[220] CAL. ELEC. CODE § 20010(c) (West 2020) (allowing victim to seek equitable relief and damages).

[221] *See Alvarez*, 567 U.S. at 725; *see also supra* notes 145-46 and accompanying text (reviewing constitutional issues with overbroad speech regulations).

[222] Actions predicated on such widespread harm would likely run afoul of the Court's standing doctrine barring adjudication of generalized grievances. *See, e.g.*, Lujan v. Defs. of Wildlife, 504 U.S. 555, 573-74 (1992) ("We have consistently held that a plaintiff raising only a generally available grievance about government—claiming only harm to his and every citizen's interest in proper application of the Constitution and laws, and seeking relief that no more directly and tangibly benefits him than it does the public at large—does not state an Article III case or controversy.").

[223] H.R. 3230 § 2.

malicious content.[224] Otherwise, a deepfake "may have already spread sufficiently to have the intended impact."[225] As noted, creating an action to enjoin platforms from publishing prohibited content or holding the online publisher liable harkens back to obstacles posed by Section 230. From a doctrinal perspective, a system for enjoining online platforms from publishing content raises concerns about government censorship and prior restraint.[226] However, a difference exists between ex ante content moderation, where online content is screened before it is published, and ex post manual content moderation, where content is either proactively or reactively flagged and reviewed after it is published.[227] A process that manually reviews and removes content, whether the function is performed voluntarily by the platform or after a court-ordered injunction, would seemingly avoid censorship concerns.

Some scholars have proposed amending Section 230 to hold online platforms liable "as the publisher or speaker" if they fail to take "reasonable steps to prevent or address unlawful uses of its services."[228] Such a change to Section 230 could subsequently hold platforms liable for failing to remove a prohibited deepfake from their platform. A legislative solution providing injunctive relief could then mirror the notice-and-takedown process provided in the Digital Millennium Copyright Act.[229] Under that process, once made aware that their platform contains material violating copyright law, the platform must remove the content and notify the alleged infringer, who can then provide counter notice to the copyright holder.[230] To remove a deepfake, the online platform could follow the same notice-and-takedown process upon receiving an injunctive order from a court.[231]

A legislative solution could also create a collaborative model, similar to the one briefly introduced in the DEEP FAKES Accountability Act,[232] to attempt to avoid the need for court-ordered injunctions. Some scholars have suggested that online platforms should take the lead in removing content themselves and work manually or automatically to flag and/or review prohibited videos, and many social media platforms are already creating algorithms to take down dubious

[224] Brown, *supra* note 162, at 14-15.

[225] *Id.* at 15.

[226] Klonick, *supra* note 204, at 1636-38.

[227] *Id.*

[228] *See* Citron & Wittes, *supra* note 209, at 419 (emphasis omitted).

[229] 17 U.S.C. § 512(g)(1) ("[A] service provider shall not be liable to any person for any claim based on the service provider's good faith disabling of access to, or removal of, material or activity claimed to be infringing . . . .").

[230] *Id.* § 512(g).

[231] Such a process assumes that the aggrieved party makes a sufficient showing of harm. *See supra* Section IV.A (describing harm requirements for First Amendment prohibitions).

[232] H.R. 3230, 116th Cong. § 6 (2019) (directing the President to make government-developed technology that reliably detects deepfakes available to Internet platforms unless doing so would harm national security interests).

content.[233] Under a collaborative model, the platform itself would serve as the adjudicator,[234] operating under a notice-and-takedown regime. Relying on cooperation from online platforms could avoid the need to amend Section 230.[235] However, although mainstream platforms like Facebook, Twitter, and YouTube may have incentives to remove prohibited content,[236] a comprehensive solution would require the unlikely scenario where all corners of the Internet comply.[237] Ultimately, an effective solution must mitigate harm by implementing a process in which malicious deepfakes can be quickly detected and removed before they metastasize online.

CONCLUSION

Without swift action, malicious deepfakes have the potential to undermine national security and election integrity. Although true ideas "tend to drive out false ones[,] [t]he problem is that the short run may be very long, that one short run follows hard upon another, and that we may become overwhelmed by the inexhaustible supply of freshly minted, often very seductive, false ideas."[238] Increasingly sophisticated technology will only make these potential threats worse. At the same time, deepfakes have numerous beneficial uses, which will likewise flourish as technology advances. The solution to this problem is a close examination of identifiable harms, followed by narrowly tailored legislation that accommodates, anticipates, and mitigates such harms without chilling expression and progress. Numerous aspects of the DEEP FAKES Accountability Act and other proposed and enacted legislation at the state and federal level provide a preliminary roadmap for creating an effective solution. However, due to deepfakes' nascent and unpredictable nature, additional considerations will inevitably arise. Crafting an appropriate response requires considering the intent

[233] *See* Brown, *supra* note 162, at 57-58.

[234] Rory Van Loo, *Federal Rules of Platform Procedure*, 88 U. CHI. L. REV. (forthcoming 2021) (manuscript at 23), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3576562 [https://perma.cc/VW8K-MJ9S] (describing how online platforms adjudicate disputes between copyright owners and alleged infringers under the DMCA).

[235] Note that private companies, such as those that host online platforms, are not subject constitutional constraints. *See, e.g.*, The Civil Rights Cases, 109 U.S. 3, 17-18 (1883) (establishing that constitutional protections granted to people through the Fourteenth Amendment only attach where government commits the violation).

[236] Klonick, *supra* note 204, at 1616-17.

[237] Though not directly relating to deepfakes, some users responded to Facebook and Twitter flagging and removing factually dubious content related to the 2020 presidential election by turning to a new (and, at that point, little-known) platform, called Parler, which they found far more welcoming to this content. *E.g.*, Kaya Yurieff, Brian Fung & Donie O'Sullivan, *Parler: Everything You Need to Know About the Banned Conservative Social Media Platform*, CNN (Jan. 10, 2021, 12:07 PM), https://www.cnn.com/2021/01/10/tech/what-is-parler/index.html [https://perma.cc/Q5YV-LPW6].

[238] Harry H. Wellington, *On Freedom of Expression*, 88 YALE L.J. 1105, 1130 (1979).

behind the creation and dissemination of a deepfake in addition to the harms posed by a deepfake’s persuasiveness and believability.