Richmond Journal of Law & Technology Volume XXVII, Issue 3

DEEPFAKES, ONLINE PLATFORMS, AND A NOVEL PROPOSAL FOR TRANSPARENCY, COLLABORATION, AND EDUCATION

Dhruva Krishna*

Cite as: Dhruva Krishna, *Deepfakes, Online Platforms, and a Novel Proposal for Transparency, Collaboration, and Education*, 27 Rich. J.L. & Tech., no. 3, 2021.

* J.D., University of California Los Angeles (2021); B.A., Carnegie Mellon University (2017). The author would like to extend thanks to the many people who have shaped this Article. I owe an enormous debt of gratitude to Professor Douglas Lichtman for sharing his mentorship, insight, and intellect throughout this process. I want to thank the faculty, staff, and my peers at UCLA Law, especially at the Ziffren Institute, who challenged me to expand my horizons and pursue my passions. To Mom, Dad, and Gopika - thank you for always inspiring me to push myself, fostering my curiosity, and reminding me that nothing is impossible with dedication and hard work. I would like to thank my friends across the world who have constantly supported me throughout all my creative, academic, and personal journeys. To Teresa – without your constant support, tremendous kindness, and magnetic energy, this Article would not have been possible. It has been an absolute pleasure to work with the Richmond JOLT Editorial Staff, whose comments, editing skills, and input have been invaluable. All errors, omissions, and mistakes are my own.

ABSTRACT

Deepfakes are manipulated media, often synthesized with machine learning, that create realistic digital impersonations, avatars, or derivative images based on pre-existing source material. Deepfakes are a source of technological innovations that can positively change our culture. However, "malicious" deepfakes pose serious threats to individuals and society at large, given their inherently upfront harm, rapid dissemination, and constant evolution that escapes an easy definition. Among privacy, technology, and legal experts, crafting policy to address malicious deepfakes has become a contentious hotspot.

This Article outlines the current gap in effective policy addressing malicious deepfakes. First, existing legal remedies are ineffective at addressing deepfake harms. Second, proposed deepfake legislation does not fare much better, by being too broad, too narrow, or utilizing impractical requirements. Third, some scholars have correctly suggested that online platforms be held responsible for deepfake regulation. However, many of these attempts, including a "reasonable steps standard," focus on stripping online platforms of important protections under Section 230 of the Communications Decency Act.

Ultimately, this Article sets forth a novel tripartite proposal to better address malicious deepfake harm while protecting technological innovation and expression. First, online platforms should provide extensive transparency disclosures to inform their users and the public more generally about their practices regarding deepfakes and manipulated media. Second, the government should collaborate with the private sector to address deepfakes. Third, both online platforms and the government should invest in public education resources about deepfakes and media literacy. This proposal best addresses the unique characteristics of malicious deepfakes, preserves technological innovation, and balances the competing values underlying powerful deepfake technology.

I. INTRODUCTION

[1] In the last four years, deepfakes have developed from an academic project to a rapidly evolving and accessible form of technology. As Part II explains, deepfakes are manipulated media, often synthesized with machine learning, that create realistic digital impersonations, avatars, or derivative images based on pre-existing source material. Like most AI developments, deepfakes can promote new artistic, technological, and scientific accomplishments.

[2] However, deepfakes have a much darker side. They can affect immense amounts of personal, economic, and reputational harm on individuals. For society, deepfakes can sew distrust, threaten democratic discourse, and raise national security concerns. Malicious deepfakes share three especially dangerous characteristics: (1) upfront harm; (2) rapid spread on social media; and (3) constant evolution that escapes an easy definition or singular targeted approach. Yet, there is debate on how best to mitigate malicious deepfake harm.

[3] Part III of this Article attempts to demonstrate that current attempts to criminalize deepfakes are ineffective. Current legal protections, including intellectual property, right of publicity, and tort protections, fail by creating unreliable claims. Recent deepfake legislation is not much better: bills are too broad, too specific, and offer ineffective solutions. Ultimately, all available legal remedies aimed at criminalizing deepfakes do not adequately respond to the three unique characteristics of malicious deepfake harm.

[4] Some scholars have suggested that online platforms should be responsible for handling deepfakes. Part IV will argue that amending Section 230 of the Communications Decency Act to hold platforms liable will impede important protections that allow for a dynamic and open internet. Specifically, adding a "reasonable steps" requirement to Section 230 immunity would impose unrealistic content moderation duties on platforms, lead to platform consolidation and confusion that could harm innovation, and would deter internet speech.

[5] Part V sets forth the following tripartite proposal of how to mitigate deepfake harms. First, online platforms should provide extensive transparency disclosures about their practices regarding deepfakes and manipulated media. Second, the federal government should collaborate with the private sector to share research, technological developments, and deepfake detection tools. Third, both online platforms and the federal government should invest in public education resources to support media literacy. Through this proposal, a better balance can be struck between preventing deepfake harm and respecting a new form of technological innovation, art, and expression.

II. DEEPFAKES: FROM FICTION TO FACT

A. What is a Deepfake?

[6] It is difficult to strictly define a deepfake.[1] At their heart, deepfakes are the "cutting-edge" trend of "increasingly realistic and convincing" digital impersonation.[2] Deepfakes can be "designed to look real….by merging, replacing, or superimposing content" onto other media.[3] These technologies can realistically mimic and alter voices, images, and human

[1] *See* James Vincent, *Why We Need a Better Definition of 'Deepfake*,' THE VERGE (May 22, 2018, 2:53 PM), https://www.theverge.com/2018/5/22/17380306/deepfake-definition-ai-manipulation-fake-news [https://perma.cc/9TJR-2KWE] (describing how legal sources have defined the term loosely and how the term "deepfake" has confusingly come to encompass a wide variety of media manipulation, including face swaps, audio manipulation, lip-synching, and more); Holly Kathleen Hall, *Deepfake Videos: When Seeing Isn't Believing*, 27 CATH. UNIV. J. L. & TECH. 51, 57 (2018) (defining deepfakes as being "created by inserting photographs into a machine-learning algorithm that puts one face on another").

[2] Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1758 (2019).

[3] Sam Shead, *Facebook to Ban 'Deepfakes*,*'* BBC NEWS (Jan. 7, 2020), https://www.bbc.com/news/technology-51018758 [https://perma.cc/NJE3-HQ3N].

expression.[4] Given the technology's rapid advancement, new methods and applications are constantly discovered.[5] Thus, the framework for what constitutes a deepfake is constantly changing. This Article defines deepfakes as manipulated media, often synthesized with machine learning, that creates realistic digital impersonations, avatars, or derivative images based on pre-existing source material.

[7] Deepfakes are often created using machine learning, neural networks, and "generative adversarial networks" (GANs).[6] Generally, neural networks utilize machine learning to process examples that "train the neural network system" and create "increasingly accurate model[s]."[7] GANs use one neural network, known as a generator, to draw on a dataset to "produce a sample that mimics the dataset."[8] A second neural network, the discriminator, then "assesses the degree to which the generator succeeded."[9] The two collaborate: the discriminator's assessments inform the generator's creations to create a realistic impersonation that "far exceeds the speed, scale, and nuance of what human reviewers could achieve."[10]

[4] *See* James Vincent, *AI Deepfakes Are Now as Simple as Typing Whatever You Want Your Subject to Say*, THE VERGE (June 10, 2019, 7:44 AM), https://www.theverge.com/2019/6/10/18659432/deepfake-ai-fakes-tech-edit-video-by-typing-new-words [https://perma.cc/723X-WL5D].

[5] *See* Chesney & Citron, *supra* note 2, at 1761–63 (explaining that multiple companies have different approaches to creating audio impersonations using sound fragments, including speech impersonation through tools such as Google DeepMind's "Wavenet" model, Baidu's "DeepVoice," and others).

[6] Chesney & Citron, *supra* note 2, at 1759–60.

[7] *Id.* at 1759.

[8] *Id.* at 1760.

[9] *Id.*

[10] *Id.*

[8] Deepfakes have become increasingly accessible and prolific. A notable example of this phenomenon occurred on Reddit in 2017.[11] A Reddit user (aptly named "deepfakes") began posting videos of celebrities superimposed on pornographic actors' bodies.[12] The user utilized machine learning algorithms, easily accessible materials, and open-source resources to create the videos.[13] These videos became widely popular and led to the development of multiple deepfake subreddits.[14]

[11] Nick Statt, *Fake Celebrity Porn Is Blowing Up on Reddit, Thanks to Artificial Intelligence*, THE VERGE (Jan. 24, 2018, 3:53 PM), https://www.theverge.com/2018/1/24/16929148/fake-celebrity-porn-ai-deepfake-face-swapping-artificial-intelligence-reddit [https://perma.cc/M7EY-8YVK].

[12] *See* Samantha Cole, *AI-Assisted Fake Porn Is Here and We're All Fucked*, MOTHERBOARD TECH BY VICE (Dec. 11, 2017, 2:18 PM), https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn [https://perma.cc/63ZL-BJV6].

[13] *Id.*

[14] *See* Samantha Cole, *Reddit Just Shut Down the Deepfakes Subreddit*, MOTHERBOARD TECH BY VICE (Feb. 7, 2018, 1:35 PM), https://www.vice.com/en/article/neqb98/reddit-shuts-down-deepfakes [https://perma.cc/6B6H-WZPH] (describing the Reddit subreddit "r/deepfakes" that approached "90,000 subscribers" and its shutdown on February 7, 2018, after allegations that it fostered involuntary pornographic content); *see, e.g.*, *Deepfakes that Are Safe for Work*, REDDIT, http://www.reddit.com/r/SFWdeepfakes/ (depicting a variety of deepfake content); *fakevideos*, REDDIT, http://www.reddit.com/r/videofakes/ [https://perma.cc/47PD-HMHA] (collecting more generally manipulated videos and deepfake content).

[9] Creators now have access to deepfake apps.[15] In January 2018, a Reddit user created an app that let anyone produce their own deepfakes.[16] The app received nearly 90,000 subscribers since its release.[17] In 2019, a Chinese company, ZAO, created a deepfake app that allowed users to place their likeness into scenes from "hundreds of movies and TV shows" with a single photo.[18] These trends demonstrate that deepfakes are sophisticated and easily distributed. Given this level of accessibility, it is important to recognize the powerful benefits and harms of deepfake use.

[15] *See, e.g.*, *Reface: Face Swap Videos*, APPLE, https://apps.apple.com/us/app/reface-face-swap-videos/id1488782587 [https://perma.cc/QW9B-C4GQ] (allowing individuals to place their likeness into popular GIFs and images); *Empowering Next-Gen Creativeness Through Technology*, BOTIKA, https://botika.io [https://perma.cc/325V-KQ6W] (allowing individuals to select still images of individuals and make them move, make personalized GIFs, and more).

[16] u/Harrumff, *Deepfake Mobile App Launch – Create Your Own High-Quality Celebrity Deepfakes in Minutes*, REDDIT, https://www.reddit.com/r/deeplearning/comments/fsmqut/deepfake_mobile_app_launch_create_your_own/ [https://perma.cc/9DGG-7MUK].

[17] *See* Cole, *supra* note 14.

[18] Jon Porter, *Another Convincing Deepfake App Goes Viral Prompting Immediate Privacy Backlash*, THE VERGE (Sept. 2, 2019, 6:32 AM), https://www.theverge.com/2019/9/2/20844338/zao-deepfake-app-movie-tv-show-face-replace-privacy-policy-concerns [https://perma.cc/D99X-CTWN] (describing the app as using cutting-edge technology that can create "convincing deepfake video[s]" with "just a single image" and the extensive privacy concerns that grant developers extensive rights and license to "all-user generated content . . . ").

B. The Benefits of Deepfakes

[10] Deepfakes are not without benefits—they allow users to create new forms of art,[19] critique, commentary, and satire.[20] Deepfake creators have used the technology to critique world leaders,[21] parody art,[22] and comment on deepfakes themselves.[23] For the entertainment industry, deepfakes may also be the future of "resurrection" technology, a means of generating new

[19] *See 3 Things You Need to Know About AI-Powered "Deep Fakes" In Art & Culture*, CUSEUM (Dec. 17, 2019), https://cuseum.com/blog/2019/12/17/3-things-you-need-to-know-about-ai-powered-deep-fakes-in-art-amp-culture [https://perma.cc/BFU7-Z9JJ] (describing various ways artists and museum are using deep fakes to enhance media projects and spaces).

[20] *See, e.g.*, Sassy Justice, *Sassy Justice with Fred Sassy (Full Episode) | From Trey Parker, Matt Stone, and Peter Serafinowicz*, YOUTUBE (Oct. 26, 2020), https://www.youtube.com/watch?v=9WfZuNceFDM [https://perma.cc/RPV2-VJV9] (explaining that the creators of "South Park," Matt Stone and Trey Parker, recently created a new web-series called "Sassy Justice" that uses deep fake technology to mock leaders and celebrities).

[21] *See, e.g.*, Ctrl Shift Face, *Better Call Trump: Money Laundering 101 [DeepFake]*, YOUTUBE (Sept. 18, 2019), https://www.youtube.com/watch?v=Ho9h0ouemWQ [https://perma.cc/BZT7-YX6M] (superimposing President Donald Trump's face and voice onto Saul Goodman from the show "Better Call Saul" to describe how to evade taxes and launder money).

[22] *See, e.g.*, Collider Extras, *George Lucas React to Star Wars: The Rise of Skywalker Final Trailer – Salty Celebrity Deepfake*, YOUTUBE (Oct. 22, 2019), https://www.youtube.com/watch?v=_MuxVqB3I7E [https://perma.cc/AG5A-B9N5] (superimposing George Lucas's face and voice onto an actor as he critiques the newest "Star Wars" entry and states that the trailer is the sound of "a thousand [Disney] executives just taking a s**t on my work").

[23] *See, e.g.*, BuzzFeedVideo, *You Won't Believe What Obama Says in This Video! [winking face emoji]*, YOUTUBE (Apr. 17, 2018), https://www.youtube.com/watch?v=cQ54GDm1eL0 [https://perma.cc/9REG-UBXS] (showing comedian Jordan Peele manipulate a deepfake of President Barack Obama to demonstrate the threats and power of deepfake technologies).

performances and works "from" deceased artists.[24] The music industry has already begun to resurrect performers, like Roy Orbison and Michael Jackson, with hologram technology.[25] In film, actors like Paul Walker, Paul Newman, Philip Seymour Hoffman, and others have been resurrected through a combination of CGI and practical effects.[26]

[11] Deepfakes will change the way we learn and interact with materials[27] by providing access to a "relatively cheap" method of producing "pedagogical" and educational content.[28] For example, students could learn

[24] *See, e.g.*, Paul Sacca, *Tupac Deepfake Raps with Snoop Dogg in New Music Video*, BROBIBLE (Mar. 8, 2020), https://brobible.com/culture/article/tupac-deepfake-snoop-dogg/ [https://perma.cc/JEZ6-XRPR].

[25] *See, e.g.*, Jefferson Graham, *The Ghost of Roy Orbison Gets a New Partner, Buddy Holly*, USA TODAY (Apr. 19, 2019, 8:54 AM), https://www.usatoday.com/story/tech/talkingtech/2019/04/19/holograms-roy-orbison-and-buddy-holly-touring-dead-live-again-shows/3473781002/ [https://perma.cc/NND4-QG94] (describing how Roy Orbison's hologram has performed for near sold-out shows in Europe); Kory Grow, *Live After Death: Inside Music's Booming New Hologram Touring Industry*, ROLLING STONE (Sept. 10, 2019, 5:06 PM), https://www.rollingstone.com/music/music-features/hologram-tours-roy-orbison-frank-zappa-whitney-houston-873399/ [https://perma.cc/PSV2-DHCV]; Zack O'Malley Greenburg, *Michael Jackson Returns to the Stage in Vegas – As a Hologram*, FORBES (May 24, 2013, 4:01 AM), https://www.forbes.com/sites/zackomalleygreenburg/2013/05/24/michael-jacksons-hologram-rocks-las-vegas-arena/#6f7ed3df3369 [https://perma.cc/JT43-JU86] (describing the Michael Jackson hologram performance in the Michael Jackson, Cirque du Soleil "ONE" show).

[26] *See* Ben Child, *From Bruce Lee to Paul Walker: How Hollywood Pulled Off Its Biggest Resurrection Acts*, GUARDIAN (June 14, 2017, 1:00 PM), https://www.theguardian.com/film/filmblog/2017/jun/14/back-from-the-dead-how-hollywood-pulled-off-its-most-unexpected-resurrection-acts [https://perma.cc/G43D-VD3J] (describing multiple resurrections throughout the history of cinema).

[27] *See, e.g.*, Chesney & Citron, *supra* note 2, at 1769–70.

[28] *Id.*

"directly" from an Abraham Lincoln avatar reading the Gettysburg Address, or Einstein teaching his theory of relativity.[29]

[12] As AI, machine learning, and technology progresses, deepfakes are likely the next frontier. AI will have significant ramifications on the healthcare industry.[30] AI and generative models could create an "entirely imaginary population of virtual patients" from existing patient data, "removing the need to share the data of real patients."[31] Researchers could "test new ways of diagnosing or monitoring disease" without risking patient privacy or data breaches.[32] The same technology being used to make fake media can be used for building speech translations and verbalizations systems, designing new civic projects, or creating cross-field tools for researchers.[33]

C. The Harms of Deepfakes

[13] Some believe that deepfakes are the beginning of an apocalyptic future—whether or not that is true—ignoring serious deepfake harm would

[29] *See id.* at 1769.

[30] *See* Geraint Rees, *Here's How Deepfake Technology Can Actually Be a Good Thing*, WORLD ECONOMIC FORUM (Nov. 25, 2019), https://www.weforum.org/agenda/2019/11/advantages-of-artificial-intelligence [https://perma.cc/B5H7-63DU].

[31] *Id.*

[32] *Id.*

[33] *See Hearing on "The National Security Challenges of Artificial Intelligence, Manipulated Media, and 'Deep Fakes'" Before the H. Permanent Select Comm. on Intel.*, 116th Cong. 3 (2019) (written testimony of Jack Clark, Policy Director, OpenAI), https://docs.house.gov/meetings/IG/IG00/20190613/109620/HHRG-116-IG00 [https://perma.cc/7YQD-9WUF] [hereinafter *Hearing on The National Security Challenges of Artificial Intelligence*].

be a mistake.[34] This Article denotes harmful deepfakes as "malicious deepfakes." Malicious deepfakes primarily affect two groups, individuals and society at large, and have three shared characteristics.

1. Commonalities of Deepfake Threats

[14] All malicious deepfakes share three characteristics. First, deepfake harm is inherently upfront.[35] Given how realistic manipulations can appear, deepfakes can reinforce stigmas, prejudices, images, and viewpoints within an audience on sight.[36] Audiovisual information is a highly persuasive form of communication because it can easily resemble the "real world."[37] These issues are more prevalent as deepfakes often include "fake news"

[34] *See, e.g.*, Dave Johnson, *How the Coming Deepfake Apocalypse Could Endanger Activism, Media, and the Truth*, MAKE CHANGE, https://makechange.aspiration.com/articles/how-the-coming-deepfake-apocalypse-could-endanger-activism-media-and-the-truth [https://perma.cc/CM66-DCS4] (discussing the potential dangers of deepfakes).

[35] *See* Steven J. Frenda et al., *False Memories of Fabricated Political Events*, 49 J. EXPERIMENTAL SOC. PSYCH. 280, 281 (2013) (stating that "visual images" have a stronger role in creating false memories and false belief).

[36] *See* Emily Thorson, Belief Echoes: The Persistent Effects of Corrected Misinformation (Jan. 1, 2013) (Ph.D. dissertation, University of Pennsylvania) (on file with Publicly Accessible Penn Dissertations) (noting that misinformation has powerful "belief echo" effects that continue to build beliefs in false information, skepticism, and challenge truth); Ullrich K.H. Ecker et. al., *Correcting False Information in Memory: Manipulating the Strength of Misinformation Encoding and its Retraction*, 18 PSYCHONOMIC BULL. & REV. 570, 571, 577 (2011) (finding that "if misinformation is encoded strongly, the level of continued influence will significantly increase" despite future corrections, *"unless* the misinformation is also retracted strongly").

[37] *See* Cristian Vaccari & Andrew Chadwick, *Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News*, SOC. MEDIA & SOC'Y, Jan.–Mar. 2020, at 2 (stating that "individuals process visual information more directly and with less effort … [and] is integrated more effectively than other types of sensory data…Misleading visuals are more likely…to generate false perceptions").

elements.[38] For example, a deepfake may include highly emotional and "unexpected" or novel content.[39] This content is especially prone to cementing notions of misinformation among watchers.[40] Continuous exposure to this misinformation builds belief in false information.[41] This is furthered by the highly realistic nature of many deepfakes, which humans have been found to discern at only 50%—as good as random guessing.[42]

[15] Second, deepfakes spread easily over social media platforms. In a landmark article about deepfakes, Citron and Chesney discuss how social media creates the perfect breeding ground for malicious deepfakes.[43] Social media platforms create "information cascades," where people start relying

[38] *See id.*; *Fake News*, DICTIONARY.COM, (2021), https://www.dictionary.com/browse/fake-news [https://perma.cc/H5NQ-TBAH] (defining "fake news" as "false news stories, often of a sensational nature, created to be widely shared or distributed for the purpose of generating revenue, or promoting or discrediting a public, political movement, company, etc.").

[39] *See, e.g.*, The Telegraph, *Jeremy Corbyn Urges Voters to Back Boris Johnson for Prime Minister in Disturbing Deepfake Video*, YOUTUBE (Nov. 12, 2019), https://www.youtube.com/watch?v=EkfnjAeHFAk [https://perma.cc/FG3A-L894] (depicting Jeremy Corbyn sponsoring Boris Johnson for Prime Minister and Boris Johnson sponsoring Jeremy Corbyn for the same position).

[40] Rachel Anne Barr, *Fake News Grabs Our Attention, Produces False Memories and Appeals to Our Emotions*, THE CONVERSATION (Nov. 17, 2019, 8:13 AM), https://theconversation.com/fake-news-grabs-our-attention-produces-false-memories-and-appeals-to-our-emotions-124842 [https://perma.cc/W8JR-89HQ].

[41] *See id.*

[42] Andreas Rössler et. al., FaceForensics: A Large-Scale Video Dataset for Forgery Detection in Human Faces 12–13 (Mar. 24, 2018) (unpublished dataset) (on file with author, University of Cornell); *see also* Donie O'Sullivan, *The Democratic Party Deepfaked its Own Chairman to Highlight 2020 Concerns*, CNN (Aug. 10, 2019, 9:58 AM), https://www.cnn.com/2019/08/09/tech/deepfake-tom-perez-dnc-defcon/index.html [https://perma.cc/6SAJ-ED3C] (noting that deepfakes tricked a room full of expert hackers at Def Con.).

[43] Chesney & Citron, *supra* note 2, at 1758.

on information they assume others have determined as true and "passing that information along."[44] More individuals pass on claims, even if they are contradictory, increasing the cascade and the credibility of the original claim despite its falsity.[45] Social media preys on human tendencies to "propagate negative and novel information."[46] Given the attention-grabbing nature of negative information, it is more clearly spread and amplified.[47] The last point the duo highlighted is that media sites amplify "filter bubbles", where individuals surround themselves with information confirming their beliefs.[48] The nature of websites whose inherent purpose is to endorse and share content has led to individuals surrounding themselves with "content from relatively homogenous groups."[49] Sites like

[44] *Id.* at 1765–66.

[45] *Id.*

[46] *Id.*

[47] *Id.* at 1767.

[48] Chesney & Citron, *supra* note 2, at 1768.

[49] *Id.*

Reddit,[50] YouTube,[51] and Facebook[52] have become hubs for deepfake creation and distribution. As more individuals consume and access information online, deepfakes could spread like wildfire.[53]

[16] The third shared characteristic of deepfakes is the constant evolution of nuanced, varied approaches. Whereas some simply manipulated media (like slowed down or sped up videos) may qualify as a deepfake, other deepfakes are more sophisticated.[54] Differing online platform policies

[50] *See, e.g.*, Cole, *supra* note 14 (noting that, before Reddit banned its deepfake subreddit, r/deepfakes, it had roughly 90,000 subscribers); *but see* r/*deepfakes*, REDDIT, https://www.reddit.com/r/deepfakes [https://perma.cc/E5BT-5XMT] (stating "r/deepfakes has been banned from Reddit") (illustrating that deepfake communities still exist on the site).

[51] *See, e.g.*, Ctrl Shift Fact, Ctrl Shift Face Uploads, YOUTUBE, https://www.youtube.com/channel/UCKpH0CKltc73e4wh0_pgL3g/videos?view=0&sort=p&flow=grid [https://perma.cc/R7QZ-4M93] (showing that one of the most notable deepfake accounts has over 429,000 subscribers and videos that have amassed over twelve million views); Ctrl Shift Fact, *Bill Hader Impersonates Arnold Schwarzenegger [Deep Fake]*, YOUTUBE (May 10, 2019), https://www.youtube.com/watch?v=bPhUhypV27w [https://perma.cc/M36B-9YBG] (showing nearly thirteen million views on a video featuring Arnold Schwarzenegger's face being superimposed on Bill Hader's face as Hader performs his impression of Schwarzenegger).

[52] *See, e.g.*, Jim Waterson, *Facebook Refuses to Delete Fake Pelosi Video Spread by Trump Supporters*, THE GUARDIAN (May 24, 2019, 3:04PM), https://www.theguardian.com/technology/2019/may/24/facebook-leaves-fake-nancy-pelosi-video-on-site [https://perma.cc/H5EQ-YZL4].

[53] *See* Chesney & Citron, *supra* note 2, at 1763–64 (discussing how the change from organizations that had a limited ability to distribute images, audio, and video to modern social media and information sites has reduced the "overall amount of gatekeeping" and democratized "access to communication to an unprecedented degree.").

[54] *Compare* Donie O'Sullivan, *Doctored Videos Shared to Make Pelosi Sound Drunk Viewed Millions of Times on Social Media*, CNN (May 24, 2019, 12:31 PM), https://www.cnn.com/2019/05/23/politics/doctored-video-pelosi/index.html [https://perma.cc/QV5Z-378D] (describing the video of Nancy Pelosi slowed down "by almost 75%" to make her appear drunk); *with* The New York Times, *Deepfakes: Is This*

demonstrate the challenges of creating a unified approach. Facebook's approach explicitly ties deepfakes to their means of creation.[55] Specifically, the technology must be a "product of artificial intelligence or machine learning."[56] In contrast, Reddit focuses on the nature of the deepfake, and considers more broadly if the deepfake is being used to impersonate and mimic users.[57] Twitter takes a hybrid approach that addresses the technical aspects of the deepfake but considers the context of the deepfake within the

Video Even Real? | NYT Opinion, YOUTUBE (Aug. 14, 2019), https://www.youtube.com/watch?v=1OqFY_2JE1c [https://perma.cc/ZQ2S-4U2U] (showing a technological expert likely using machine learning and AI programs to create a highly realistic impersonation of pop singer Adele). *See also* Britt Paris & Joan Donovan, *Deepfakes and Cheapfakes*: *The Manipulation of Audio and Visual Evidence*, DATA & SOCIETY at 11–16, https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1.pdf [https://perma.cc/7N2K-6UJ7] (describing a spectrum of deepfake content, with some requiring less expertise and fewer technical resources, like "in-camera effects," while others utilize more expertise and technical resources required, like computer neural networks).

[55] *Manipulated Media*, *Community Standards*, FACEBOOK, https://www.facebook.com/communitystandards/manipulated_media [https://perma.cc/F3TD-8S2U].

[56] *Id.* ("Video[] that ha[s] been edited or synthesized, beyond adjustments for clarity or quality, in ways that are not apparent to an average person…AND *is the product of artificial intelligence or machine learning*, including deep learning techniques (e.g. a technical deepfake) that merges, combines, replaces, and/or superimposes content onto a video, creating a video that appears authentic") (emphasis added).

[57] *Do Not Impersonate an Individual or Entity*, *Account and Community Restrictions*, REDDIT (July 2020), https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/do-not-impersonate-individual-or [https://perma.cc/PM2E-ZCC2]; *see also Do Not Post Involuntary Pornography*, *Account and Community Restrictions*, REDDIT (July 2020), https://www.reddithelp.com/hc/en-us/articles/360043513411 [https://perma.cc/RUK5-6P9P] ("Reddit prohibits the dissemination of images or video depicting any person in a state of nudity or engaged in any act of sexual conduct apparently created or posted without their permission, including depictions that have been faked.").

platform's larger ecosystem.[58] Twitter and Facebook explicitly consider whether the content would impact public safety or cause serious harm in their policies,[59] whereas that may be an implicit value to Reddit.[60] These examples of differing policies show that that even for the online platforms on the frontline of deepfake issues, defining deepfakes is a debated and contested judgement.

2. Harm Against Individuals

[17] The shared characteristics of malicious deepfakes mean that deepfakes pose serious personal harm for individuals. The wide sharing of deepfake technology and online communities allows users to easily create pornographic images without people's consent.[61] Celebrities have

[58] *See* Yoel Roth & Ashita Achuthan, *Building Rules in Public: Our Approach to Synthetic & Manipulated Media*, TWITTER (Feb. 4, 2020), https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html [https://perma.cc/CV8Z-T6QW].

[59] *See id.*; Monika Bickert, *Enforcing Against Manipulated Media*, FACEBOOK (Jan. 6, 2020), https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/ [https://perma.cc/CMY6-XQG7]; *Violence and Incitement*, *Community Standards*, FACEBOOK, https://www.facebook.com/communitystandards/credible_violence [https://perma.cc/6WXQ-48XJ] ("We remove content, disable accounts, and work with law enforcement when we believe there is a genuine risk of physical harm or direct threats to public safety.").

[60] *See Do Not Post Violent Content*, *Account and Community Restrictions*, REDDIT (June 2020), https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/do-not-post-violent-content [https://perma.cc/85JC-WMMS] ("Do not post content that encourages, glorifies, incites, or calls for violence or physical harm against an individual or a group of people…or encourages the abuse of animals.").

[61] *See, e.g.*, James Vincent, *New AI Deepfake App Creates Nude Images of Women in Seconds*, THE VERGE (June 27, 2019, 6:23 AM), https://www.theverge.com/2019/6/27/18760896/deepfake-nude-ai-app-women-deepnude-non-consensual-pornography [https://perma.cc/JS55-W3PM]; Samantha Cole, *This Horrifying App Undresses a Photo of Any Woman with a Single Click*, MOTHERBOARD TECH BY VICE (June 26, 2019, 5:49 PM), https://www.vice.com/en_us/article/kzm59x/deepnude-app-creates-fake-nudes-of-any-

expressed their horror at the distribution of highly realistic pornographic videos featuring their likenesses.[62] These videos exploit private individuals too: deepfake applications like DeepNude allowed users to create pornographic images of any individual.[63] All victims of pornographic deepfake videos suffer from profound psychological damage as they are reduced "to sex objects."[64] Individuals may feel threatened, as deepfakes can lead to the loss of romantic opportunities, the cancellation of business opportunities, and the loss of personal relationships.[65] Given how easily content and "viral" matter can be transmitted today,[66] deepfake videos will

woman [https://perma.cc/65AF-D6QW] (reporting that the DeepNude app, created using an open-source algorithm developed by University of California, Berkeley researchers, was available in a free and $50 paid version, allowing users to create nude images of women, with easily removable watermarks indicating that the photo was fake).

[62] *See, e.g.*, Statt, *supra* note 11; Sean Hollister, *Scarlett Johansson Slams Deepfakes, Says She Can't Stop the Internet from Pasting Her Face on Porn*, THE VERGE (Dec. 31, 2018, 5:30 PM), https://www.theverge.com/2018/12/31/18163351/scarlett-johansson-slams-deepfakes-internet-lost-cause [https://perma.cc/52BR-PY4E].

[63] *See* Samantha Cole, *This Horrifying App Undresses a Photo of Any Woman with a Single Click*, MOTHERBOARD TECH BY VICE (June 26, 2019, 5:49 PM), https://www.vice.com/en_us/article/kzm59x/deepnude-app-creates-fake-nudes-of-any-woman [https://perma.cc/65AF-D6QW] (explaining that DeepNude created deepfake nude images of any individual, often with better results using "high resolution images").

[64] Chesney & Citron, *supra* note 2, at 1773.

[65] *Id.* at 1774; *Cf.* Samantha Cole & Emanuel Maiberg, *Pornhub Doesn't Care*, MOTHERBOARD, TECH BY VICE (Feb. 6, 2020, 9:00 AM), https://www.vice.com/en_us/article/9393zp/how-pornhub-moderation-works-girls-do-porn [https://perma.cc/C66S-2Z9C] (discussing how recent litigation surrounding "Girls Do Porn," although not specifically deepfakes, shows how unauthorized pornographic videos can have serious ramifications on the victims' lives despite the "official site [being] shut down" as "hundreds of 'Girls Do Porn' videos are easy to find" on sites like PornHub).

[66] *See, e.g.*, Chris Meserole, *How Misinformation Spreads on Social Media–And What to Do About It*, BROOKINGS (May 9, 2018), https://www.brookings.edu/blog/order-from-chaos/2018/05/09/how-misinformation-spreads-on-social-media-and-what-to-do-about-it/

harm the reputations of individuals long before the videos are disavowed, removed, or otherwise addressed.

[18] Deepfakes may inflict other types of individual harm. In 2019, an energy company's chief executive was deceived into wiring about $220,000 to a Hungarian supplier.[67] The thief utilized AI technology and deepfake software to mimic the voice, tonality, punctuation, and German accent of the executive's boss to execute the transfer.[68] Given how quickly deepfake technology is progressing, it is feasible to imagine a world where such AI-driven crimes become increasingly common.[69] Of these crimes, deepfakes were rated the most dangerous as the "most-concerning," with potential applications to "exploit people's implicit trust" to "gain access to funds," "request access to secure systems," or create larger social harms.[70]

3. Harm Against Society

[19] Deepfakes pose a profound threat to society. The spread of altered videos of Speaker Nancy Pelosi in 2019 exemplifies how deepfakes sew

[https://perma.cc/7KK6-FQQ8] (describing how an inaccurate tweet describing the perpetuator of a terrorist act as "Middle-Eastern" was retweeted almost 1400 times in about five hours, whereas the later clarification that identified the perpetuator as white had under 200 retweets in the same time period).

[67] Nick Statt, *Thieves Are Now Using AI Deepfakes to Trick Companies into Sending Them Money*, THE VERGE (Sept. 5, 2019, 1:14 PM), https://www.theverge.com/2019/9/5/20851248/deepfakes-ai-fake-audio-phone-calls-thieves-trick-companies-stealing-money [https://perma.cc/QSP4-NQ9T].

[68] *Id.*

[69] *See generally* M. Caldwell et al., *AI-Enabled Future Crime*, CRIME SCI. 9 (2020), https://doi.org/10.1186/s40163-020-00123-8 [https://perma.cc/F3SB-GXFA] (discussing the applications of "artificial intelligence and related technologies in the perpetration of crime" in areas like impersonation, weapons, and more, with deepfakes being rated the most dangerous, with potential applications to "exploit people's implicit trust" to "gain access to funds," "request access to secure systems," or create larger societal harms).

[70] *Id.*

public distrust.[71] The videos were circulated on major platforms such as Facebook and Twitter.[72] One video, deemed a "low-tech fake" because it slowed down an existing video of Pelosi without AI,[73] depicted Pelosi slurring her words and appearing drunk.[74] President Trump retweeted the deepfake video of Pelosi,[75] and it was viewed millions of times online.[76] Facebook refused to remove the video, stating it had not violated its community guidelines.[77] However, it took seemingly inconsequential steps

[71] Sarah Mervosh, *Distorted Videos of Nancy Pelosi, Spread on Facebook and Twitter, Helped by Trump*, N.Y. TIMES (May 24, 2019), https://www.nytimes.com/2019/05/24/us/politics/pelosi-doctored-video.html [https://perma.cc/2BYU-YR8X]; Donie O'Sullivan, *Doctored Videos Shared to Make Pelosi Sound Drunk Viewed Millions of Times on Social Media*, CNN (May 24, 2019, 12:31 PM), https://www.cnn.com/2019/05/23/politics/doctored-video-pelosi/index.html [https://perma.cc/QV5Z-378D].

[72] Sarah Mervosh, *supra* note 71.

[73] Olivia Beavers, *House Intel to Take First Major Deep Dive into Threat of 'Deepfakes'*, THE HILL (June 13, 2019, 6:00 AM), https://thehill.com/homenews/house/448278-house-intel-to-take-first-major-deep-dive-into-threat-of-deepfakes [https://perma.cc/8789-V2YX] (noting that "despite the 'simple manipulation,' the 'low-tech fake' demonstrated how dangerous and believable manipulated video content can be.).

[74] Sarah Mervosh, *supra* note 71.

[75] Donald Trump (@realDonald Trump), *"Pelosi Stammers Through News Conference"*, TWITTER (May 23, 2019, 9:09 PM), https://web.archive.org/web/20190524010938if_/https://twitter.com/realDonaldTrump/status/1131728912835383300 [https://perma.cc/XBH7-N549].

[76] Sue Halpern, *Facebook's False Standards For Not Removing A Fake Nancy Pelosi Video*, THE NEW YORKER (May 28, 2019), https://www.newyorker.com/tech/annals-of-technology/facebooks-false-standards-for-not-removing-a-fake-nancy-pelosi-video [https://perma.cc/UHV8-YFH6].

[77] *Id.* (discussing how "Facebook refused to remove the Pelosi video because…it does not violate the company's community standards, even though it is demonstrably false."); *see also* Emily Stewart, *A Fake Viral Video Makes Nancy Pelosi Look Drunk. Facebook Won't Take it Down*, VOX (May 24, 2019, 3:50 PM), https://www.vox.com/recode/2019/5/24/18638822/nancy-pelosi-doctored-video-drunk-facebook-trump [https://perma.cc/BC52-6ETV].

to prevent its spread by "reducing its distribution" and providing video "context."[78] In contrast, YouTube promptly removed the video, but it was still found in different iterations on the platform.[79] The video was available on President Trump's Twitter account, before it was suspended in 2021.[80]

[20] The Pelosi video sewed immense partisan hostility and debate.[81] One could imagine a more realistic video creating more discord. Manipulated media of law enforcement, a government figure, or even a celebrity would create distrust of figures, institutions, and organizations.[82]

[21] The existence of deepfakes calls into question the legitimacy of democratic discourse and processes. Both parties were concerned that a malicious deepfake would severely disrupt the 2020 election.[83] These

[78] Stewart, *supra* note 77.

[79] Makena Kelly, *Trump Tests Disinformation Policies with New Pelosi Video*, THE VERGE (Feb. 7, 2020, 2:28 PM), https://www.theverge.com/2020/2/7/21128317/nancy-pelosi-donald-trump-disinformation-policy-video-state-of-the-untion [https://perma.cc/T6JY-CPW6].

[80] Trump, *supra* note 75; Gabrielle Chung, *Donald Trump's Twitter Account Permanently Suspended 'Due to the Risk of Further Incitement of Violence'*, PEOPLE (Jan 8, 2021), https://people.com/politics/donald-trump-twitter-account-permanently-suspended/ [https://perma.cc/ET2K-JYLU].

[81] *See* Daniel Funke, *Why False Claims About Nancy Pelosi Being Drunk Keep Going Viral – Even Though She Doesn't Drink*, POYNTER (Aug. 4, 2020), https://www.poynter.org/fact-checking/2020/why-false-claims-about-nancy-pelosi-being-drunk-keep-going-viral-she-doesnt-drink/ [https://perma.cc/9HBY-HCMY].

[82] *See* Riana Pfefferkorn, *Too Good to Be True? "Deepfakes" Pose a New Challenge for Trial Courts*, NW LAWYER. Sept. 2019, at 23 (stating that some have argued that deepfakes could even change standards of evidence in trial courts).

[83] *See* DANIEL R. COATS, STATEMENT FOR THE RECORD: WORLDWIDE THREAT ASSESSMENT OF THE US INTELLIGENCE COMMUNITY, Senate Select Committee on Intelligence 7 (2019), https://www.dni.gov/files/ODNI/documents/2019-ATA-SFR---SSCI.pdf [https://perma.cc/J288-VXLL] ("Adversaries and strategic competitors probably will attempt to use deep fakes . . . to augment influence campaigns directed

concerns were amplified by interference in the 2016 United States presidential election and the cyber-espionage campaign against French President Emmanuel Macron in 2017.[84]

[22] Deepfakes pose a threat to diplomacy, public safety, and global security.[85] Deepfakes of officials may weaken reputations, force a government into conflict, and create domestic unrest at home.[86] A deepfake of a government official stating there was a terrorist attack, disease outbreak, or chemical accident could sow panic. State and non-state actors could make deepfakes to create disruptions.[87] Given these threats, malicious deepfakes require a response.

III. CRIMINALIZING DEEPFAKES IS INEFFECTIVE AND DANGEROUS

[23] One common response by policymakers is to criminalize deepfakes.[88] This criminalization is done in two ways. One is by using current legal protections to encompass deepfake threats. The other is the proposal and enactment of new statutes that specifically address deepfakes. Neither method effectively addresses deepfakes and endangers a valuable form of technology.

against the United States . . . "); Cristiano Lima, *'Nightmarish': Lawmakers Brace for Swarm of 2020 Deepfakes*, POLITICO (June 13, 2019, 1:43 PM), https://www.politico.com/story/2019/06/13/facebook-deep-fakes-2020-1527268 [https://perma.cc/X7RK-ZPNN] (describing how both GOP and Democrat candidates have engaged in protocols to monitor for manipulated media and deep forgeries targeting their candidates).

[84] Chesney & Citron, *supra* note 2, at 1778.

[85] *Id.* at 1782–84.

[86] *See id.* at 1782.

[87] *See id.*

[88] *Id.* at 1803.

A. Current Legal Protections Do Not Adequately Respond to Deepfakes

[24] There are currently several legal remedies that may be used to criminalize deepfakes. Intellectual property claims are an example. Copyright protects "original works of authorship fixed in any tangible medium of expression."[89] Trademarks prevent the use of a mark "on or in connection with goods and/or services in a manner that is likely to cause confusion, deception, or mistake about the source of the goods and/or services."[90] A deepfake creator may take content from several movies to create their material.[91] In these cases, the studios and content owners could pursue copyright or trademark claims against creators. This may lead to takedown notices or using other avenues of recourse.[92]

[25] However, intellectual property remedies are inadequate for most deepfakes. Fair use considerations will constrain enforcement.[93] For example, an innocuous use of an actor's face in a video may be found to be "transformative use" under copyright doctrine,[94] but these same protections may extend to deepfakes that create societal harm. Deepfake creators who use underlying source material from news agencies or of government

[89] Copyright Act of 1976, 17 U.S.C. §102.

[90] *About Trademark Infringement*, USPTO (June 8, 2018. 2:03 PM), https://www.uspto.gov/page/about-trademark-infringement [https://perma.cc/7QKD-5S84].

[91] *See* Porter, *supra* note 18.

[92] *See* YouTube Creators, *Copyright Takedowns & Content ID – Copyright on YouTube*, YOUTUBE (Oct. 12, 2020), https://www.youtube.com/watch?v=4qfV0PRsCrs&feature=emb_logo [https://perma.cc/C56L-EFC9].

[93] Copyright Act of 1976, 17 U.S.C. §107.

[94] *See* Michael J. Madison, *A Pattern-Oriented Approach to Fair Use*, 45 WM. & MARY L. REV. 1525, 1670 (2004).

officials blur the line between commentary and malicious activity. Trademark law may have little application if consumers are insufficiently confused or if the video does not sell a product.[95] For example, a court may find that few individuals will think that a video superimposing Nicholas Cage's face onto a character in Sesame Street would believe the video is sponsored by, or originates from, Nicholas Cage.[96]

[26] The right of publicity may also provide some limited remedies. The right of publicity is a state law that "provides a basis to control the unwanted dissemination of one's name and likeness, and other indicia of identity for another's advantage."[97] Rights of publicity do not exist federally,[98] but may be recognized in statute and in common law,[99] and may grant postmortem

[95] *See, e.g.*, TwinkieMan, *Nicholas Cage Sesame Street [Deepfake]*, YOUTUBE (Aug. 9, 2019), https://www.youtube.com/watch?v=BN-uBfAq9jY [https://perma.cc/8M96-9C4E] (depicting a video superimposing Nicholas Cage's face onto a character in Sesame Street).

[96] *See id.*

[97] *See* Alden Hunt, *Jennifer Rothman '91 Explains the Right to Privacy*, PRINCETON ALUMNI WEEKLY (June 14, 2018), https://paw.princeton.edu/article/jennifer-rothman-91-explains-right-publicity#:~:text=Instead%2C%20the%20right%20of%20publicity,and%20not%20a%20uniform%20one.&text=At%20its%20broadest%2C%20the%20right,of%20identity%20for%20another's%20advantage.%E2%80%9D [https://perma.cc/YQ4V-SRQK]; *see also* Samuel D. Warren & Louis D. Brandeis, *The Right To Privacy*, 4 HARV. L. REV. 193, 193, 196 (1890) (noting an individual's legally recognized right to privacy); Zacchini v. Scripps-Howard Broad. Co., 433 U.S. 562, 573 (1977) (stating that the right of publicity explicitly protects the "commercial benefit" of the entertainer).

[98] JENNIFER ROTHMAN, THE RIGHT OF PUBLICITY: PRIVACY REIMAGINED FOR A PUBLIC WORLD 3 (2018).

[99] *See The Law*, ROTHMAN'S ROADMAP TO THE RIGHT OF PUBLICITY, https://www.rightofpublicityroadmap.com/law [https://perma.cc/Y3BS-8PD5] (providing a thorough overview of each state's various right of publicity laws).

protections for various lengths of time.[100] A right of publicity claim would work best if a deepfake creator was using one's likeness for profit or to commercialize their own products.[101]

[27] However, the right of publicity would do little to mitigate deepfake harms. First, victims may have a difficult time pursuing right of publicity claims if they did not create economic benefit in their likeness.[102] Second, the case law surrounding the right of publicity is incohesive and would not be adequate for solving deepfake harms federally—including California, there are at least five balancing approaches to the right of publicity nationally.[103] These various approaches have "led to bizarre and conflicting outcomes in cases with similar facts."[104] Third, these protections may be difficult if deepfakes are not being used for commercial uses. If a malicious actor merely posted a deepfake of a politician on multiple platforms without attempting to monetize the video, there seems to be no "hook" for enforcement.

---

[100] *See* ROTHMAN, *supra* note 98, at 97–98; *see also* Milton H. Greene Archives, Inc., v. Marilyn Monroe LLC, 692 F.3d 983, 986 (9th Cir. 2012) (finding that because Marilyn Monroe was domiciled in New York at her death, she could not exercise California's postmortem right of publicity).

[101] *See, e.g.*, White v. Samsung Elecs. Am., Inc., 971 F.2d 1395, 1399 (9th Cir. 1992) (finding that a robot that merely resembled TV host Vanna White by being "dressed in a wig, gown, and jewelry," could present a plausible right of publicity claim given White's "sole right to exploit" her "celebrity value"); *see* Waits v. Frito-Lay, Inc., 978 F.2d 1093, 1112 (9th Cir. 1992) (finding the defendant liable for violating the plaintiff's right of publicity by hiring a sound-alike to imitate his voice in a commercial).

[102] *See* Sarver v. Chartier, 813 F.3d 891, 905–06 (9th Cir. 2016) (finding that because the defendant was a "private person" and never "exploited the economic value of any performance or persona he had worked to develop," the state had "no interest in giving Sarver an economic incentive" and could not pursue a right of publicity claim).

[103] *See* ROTHMAN, *supra* note 98, at 145–147 (discussing the five distinct approaches: *ad hoc* balancing approach, the *transformative-work* test, the *transformative-use* test, the *relatedness* test, and the *predominant-purpose* test).

[104] *See id.* at 147.

[28] Tort protections may address deepfakes.[105] Two potential tort remedies are false light claims and defamation.[106] False light claims "commonly address photo manipulation, embellishment, and distortion, as well as deceptive uses of non-manipulated photos."[107] Defamation claims are similar, but differ given the claimed injury.[108] Defamation "compensates for damage to reputation," whereas "false light compensates for being subject to offensiveness."[109] Individual plaintiffs would likely use these protections to show that a defendant created a deepfake that gave a false or misleading impression of the plaintiff that damaged their reputation or caused great offense.[110]

[29] These tort protections would fail. For public figures, an "actual malice" requirement is often required for both false light and defamation claims.[111] This creates difficult hurdles for celebrities portrayed in pornographic videos.[112] Tort remedies may also inadequately address

[105] *See* Chesney & Citron, *supra* note 2, at 1793–94.

[106] *See id.*

[107] David Greene, *We Don't Need New Laws for Faked Videos, We Already Have Them*, ELECTRONIC FRONTIER FOUND. (Feb. 13, 2018), https://www.eff.org/deeplinks/2018/02/we-dont-need-new-laws-faked-videos-we-already-have-them. [https://perma.cc/VAK4-VVZE].

[108] *See id.*

[109] *Id.*

[110] *See id.*

[111] *See id.*

[112] *See* Chesney & Citron, *supra* note 2, at 1793 ("Public officials and public figures are subject to a higher requirement of showing clear and convincing evidence of actual malice"); David Singer & Camila Connolly, *How Hollywood Can (and Can't) Fight Back Against Deepfake Videos (Guest Column)*, THE HOLLYWOOD REPORTER (Sept. 7, 2019), https://www.hollywoodreporter.com/thr-esq/how-hollywood-can-can-t-fight-back-deepfake-videos-guest-column-1237685 [https://perma.cc/4PYP-LXXS] (stating that,

societal harms. For manipulated videos of a world leader, the higher bar for actual malice could make claims untenable.[113] If a manipulated deepfake video portrayed an attack on a town of faked residents, the individuals whose images were used may not be sufficiently harmed under these tort protections.

B. New Legislative Proposals to Criminalize Deepfakes Are Also Ineffective

[30] Given deepfake threats and the failure of current legal remedies, bills have been proposed and enacted to criminalize deepfakes. However, these bills fail to address deepfake harms.

1. Overly Broad Proposals Fail to Give Guidance and Lack Substance

[31] One class of bills seeks to remedy deepfake threats through overly broad legislation. These bills have been proposed or passed in several states including New York,[114] Massachusetts,[115] and federally.[116] One federal bill,

"[i]n most states, existing defamation, right of publicity and invasion of privacy laws will not reach deepfakes . . . ").

[113] *See* Michael Scott Henderson, *Applying Tort Law to Fabricated Digital Content*, 5 UTAH L. REV. 1145, 1167 (2020) (finding that the "actual malice" requirement would make enforcement of defamation and false light claims by public figures difficult).

[114] *See* Assemb. B. A8155B, 2017–18 Reg. Sess., (NY 2017), https://www.nysenate.gov/legislation/bills/2017/a8155 [https://perma.cc/7VQN-PNZS] (attempting to provide New York with a postmortem right of publicity that also protects against digital replicas that reproduce a "living or deceased individual's likeness or voice").

[115] *See* H.R. 3366, 191 General Ct. (Ma. 2019), https://malegislature.gov/Bills/191/H3366. [https://perma.cc/Y9SL-AGD5].

[116] Malicious Deep Fake Prohibition Act of 2018, S. 3805, 115th Cong. (2018), https://www.congress.gov/bill/115th-congress/senatebill/3805/text?q=%7B%22search%22%3A%5B%22deep+fake%22%5D%7D&r=34&s=4 [https://perma.cc/U5JM-ASFU].

the "Malicious Deep Fake Prohibition Act of 2018," proposed by Senator Ben Sasse is particularly egregious."[117]

[32] Senator Sasse's bill defines deepfakes as "an audiovisual record created or altered . . . [so] that the record would falsely appear to a reasonable observer to be . . . [an individual's] actual speech or conduct . . . ."[118] An audiovisual record is defined as, "any audio or visual media in an electronic format."[119]

[33] This language is poorly written and overinclusive. Imagine a musician uploads a remix of a sound recording to YouTube.[120] To the "reasonable observer," this unauthorized remix seems to be the original artist's "authentic record." Although this remix may be illegal, it would seem strange to consider it a "deepfake." This is far from the pornographic, political, or even manipulated "cheapfake" videos that cause serious harm.[121]

[34] An overinclusive definition will lead to overcriminalization. Senator Sasse's bill criminalizes creating or distributing any deepfake that may

[117] *Id.*

[118] *Id.* at §1041(a).

[119] *Id.*

[120] *See, e.g.*, Bistro Music, *The Beach Boys- Wouldn't it be nice (Chipper Fresco Remix) [CHILL]*, YOUTUBE (Jun. 7, 2017), https://www.youtube.com/watch?v=IGrE_5FHz2I [https://perma.cc/LL5L-JW6P].

[121] *See* Britt Paris & Joan Donovan, *Deepfakes and Cheapfakes*: *The Manipulation of Audio and Visual Evidence*, DATA & SOCIETY at 11–16, https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1.pdf [https://perma.cc/P6JN-J5UJ] (describing a spectrum of deepfake content, with some requiring less expertise and fewer technical resources, like "in-camera effects," while others utilize more expertise and technical resources required, like computer neural networks).

facilitate "criminal or tortious conduct."[122] Yet, it is already illegal to commit or facilitate a crime under federal and multiple state laws.[123] This bill simply adds a "federal criminal law hammer to conduct that is already prohibited."[124] Although this redundancy may be attractive to deepfake critics who want multiple avenues of liability, it will have overly expansive effects. Suppose our musician throws a party and plays their music very loudly. If the musician received a noise complaint or violated sound ordinances, would their conduct now be a federal crime? Under Senator Sasse's bill, it may well be.[125] This seems to be far from what the "drafters of the bill intended" to be in the bill's scope.[126] This is serious given the penalties in Sasse's bill—potentially ten years in prison.[127] Ultimately, Sasse's bill, and others like it, would lead to frivolous litigation and deter innovation for beneficial deepfake uses.

[122] Malicious Deep Fake Prohibition Act of 2018, S. 3805, 115th Cong. §1041(b) (2018).

[123] Orin S. Kerr, *Should Congress Pass A "Deep Fakes" Law?*, THE VOLOKH CONSPIRACY (Jan. 31, 2019), https://reason.com/2019/01/31/should-congress-pass-a-deep-fakes-law/ [https://perma.cc/V7FD-JUGD].

[124] *Id.*

[125] *See id.* (coming to a similar conclusion because the individual hosting the party would be "distributing copies of a deepfake" and doing so with "the intent to facilitate conduct that is a tortious nuisance under state law" by hosting a loud party).

[126] Kerr, *supra* note 123.

[127] *See* Malicious Deep Fake Prohibition Act of 2018, S. 3805, 115th Cong. § 1041(c)(2).

2. Proposals that Ban Specific Deepfakes Are Too Narrow

[35] A second class of legislation bans deepfakes in specific contexts. These types of bills have been proposed or passed in California,[128] Texas,[129] Virginia,[130] and federally.[131] However, these bills' specificity make them inadequate addresses of deepfakes.

[36] California Bill AB-602 was recently approved and specifically targets deepfakes in pornographic uses.[132] The bill creates two causes of action.[133] First, against a deepfake creator who "creates and intentionally discloses sexually explicit material" without the depicted individual's

[128] *See* Depiction of individual using digital or electronic technology: sexually explicit material: cause of action, B. 602, 2019 Gen. Assemb., Reg. Sess. §1708.86 (Ca. 2019), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB602 [https://perma.cc/XA7B-PH2V].

[129] *See* S. 751, 2019 Leg., 88th Sess. (Tex. 2019), https://legiscan.com/TX/text/SB751/id/2027638 [https://perma.cc/7RBJ-67SP] (creating a "criminal offense for fabricating a deceptive video with intent to influence the outcome of an election").

[130] *See* Unlawful creation of image of another; penalty, VA. CODE ANN. §18.2-386.1 (2019), https://law.lis.virginia.gov/vacode/title18.2/chapter8/section18.2-386.1/ [https://perma.cc/4DSF-MKVM] (creating a criminal offense for pornographic deepfakes by creating "videographic or still image[s] by any means").

[131] *See, e.g.*, Deepfakes in Federal Elections Prohibition Act, H.R. 6088, 116th Cong. (2020), https://www.congress.gov/bill/116th-congress/house-bill/6088/titles [https://perma.cc/H79R-F57L] (wanting to "amend the Federal Election Campaign Act of 1971 to prohibit the distribution of materially deceptive audio or visual media prior to an election for Federal office . . . ").

[132] Depiction of individual using digital or electronic technology: sexually explicit material: cause of action, §1708.86, B. 602, Assemb., Reg. Sess. (Ca. 2019), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB602 [https://perma.cc/S3QR-CPVS].

[133] *Id.* at §1708.86(b).

consent.[134] Second, the bill creates a cause of action against a deepfake distributor who "intentionally discloses" this sexually explicit material, knowing the depicted individual did not consent to its creation.[135] Ultimately, a successful plaintiff could seek maximum economic damages of $150,000.[136]

[37] AB-602 more narrowly limits what deepfakes it encompasses. The bill limits "nude" to mean specific body parts like "visible genitals."[137] "Sexually explicit material" is limited to "any portion of an audiovisual work" that depicts the individual "performing in the nude or appearing to engage in, or being subjected to, sexual conduct."[138]

[38] However, the bill is not without its flaws.[139] The bill's biggest strength, its specificity, is also its biggest weakness. The bill does little to remedy other forms of deepfake harm. If an individual's likeness is used in

[134] *Id.* at §1708.86(b)(1).

[135] *Id.* at §1708.86(b)(2)*; see id.* at §1708.86(c) (stating exceptions, specifically if the content is "a matter of legitimate public concern," "a work of political or newsworthy value or similar work," or "commentary, criticism or disclosure that is otherwise protected by the California Constitution or the United States Constitution").

[136] *See* Assemb. B. 602, Ch. 491 § (e)(1)(B)(ii) (Cal. 2019).

[137] *Id.* § (1)(a)(10).

[138] *Id.* § (1)(a)(14). *See also, id.* § (1)(a)(13) (defining sexual conduct to include a specific list of activities including masturbation, sexual intercourse regardless of sex or gender or between humans and animals, sexual penetration, the transfer of semen, or sadomasochistic abuse).

[139] *See* Kamran Salour & Veronica Reynolds, *If Signed by Governor, California Bill AB-602 Will Provide Private Right of Action for Victims of Sexually Explicit 'Deepfakes,'* BAKER HOSTETLER DATA COUNSEL (Sept. 26, 2019), https://www.dataprivacymonitor.com/state-legislation/if-signed-by-governor-california-bill-ab-602-will-provide-private-right-of-action-for-victims-of-sexually-explicit-deepfakes/ [https://perma.cc/3L7E-5VYH] (describing the limitations of the bill, including potential preemption, lack of postmortem rights, and vague notions of consent).

a video portraying them robbing a store or engaging in a racist tirade, these uses would not be encompassed. Expecting state or federal legislature to continuously pass specific-act-legislation may lead to overenforcement or overinclusive bills.

3. Proposals that Mandate Specific Requirements Fail to Recognize Technological Change

[39] A third set of bills seeks to mitigate deepfake harm by mandating specific requirements. Such bills have been proposed and passed in California and at the federal level.[140] One of the most discussed bills in this class is Yvette Clarke's DEEPFAKES Accountability Act of 2019.[141]

[40] Congresswoman Clarke's bill proscribes mandatory disclosures on seemingly all manipulated media.[142] For manipulated media with a "moving visual element," the bill mandates an "embedded digital watermark" identifying the record as altered.[143] For those containing a "visual element" there must be "an unobscured written statement . . . throughout the duration of the visual element" stating there are "altered visual elements" with a "concise description" of the alteration.[144] For those containing an audio element, there must be "a clearly articulated verbal statement that identifies the record" as being altered, with a "concise description" of the alteration."[145] Failure to include these disclosures could lead to fines,

[140] *See* Assemb. B. 730, Ch. 493 § 4(b)(3)(A–B) (Cal. 2019) (adding to Section 20010 of the Elections Code, relating to Elections); DEEP FAKES Accountability Act, H.R. 3230, 116th Cong. (2019).

[141] *See* DEEP FAKES Accountability Act, H.R. 3230, 116th Cong. (2019).

[142] *Id*. § 1041.

[143] *Id.* § 1041(b).

[144] *Id.* § 1041(d).

[145] *Id.* § 1041(e).

damages, and federal prosecution.[146] Although the bill grants the Attorney General power to issue waivers for manipulated media that may require a watermark if they are protected by the First Amendment, this exclusion still seems likely to deter speech.[147]

[41] The bill also ignores technological realities. Users who are willing to watermark their videos are those who use deepfakes for innocuous reasons.[148] Malicious deepfake creators could intentionally avoid these disclosures.[149] As videos are spread online, a lack of watermarks would only aid malicious deepfake creators to make deceptive videos seem authentic. Removing watermarks, a relatively easy process, on all deepfake content would create further confusion.[150] Although Congresswoman Clarke's bill creates liability for removing such watermarks,[151] finding the remover may be impossible. Rather than being flexible, legislative proposals (like Clarke's) that are too specific will not address the technological realities of deepfakes.

C. Any Attempt to Criminalize Deepfakes Directly Is Flawed

[42] Criminalizing deepfakes is insufficient because post facto criminalization cannot undo deepfake harm. Imposing creator liability is

[146] *See id*. at §1041(f).

[147] *See* H.R. 3230, at §1041(k)(2).

[148] *See* Devin Coldewey, *DEEPFAKES Accountability Act Would Impose Unenforceable Rules- But It's a Start*, TECHCRUNCH (June 13, 2019, 3:25 PM), https://techcrunch.com/2019/06/13/deepfakes-accountability-act-would-impose-unenforceable-rules-but-its-a-start/ [https://perma.cc/2AUC-AU2].

[149] *See id.*

[150] *See id.*

[151] *See* Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019, H.R. 3230, 116th Cong. §1041(f)(1)(B) (2019).

unlikely to create deterrence or successful suits.[152] Instead, suing deepfake creators may be impossible.[153] For example, imagine a deepfake creator, X, creates a pornographic deepfake of individual Y. The video is distributed and reshared dozens, hundreds, and possibly thousands of times online. Once Y receives notice of the video, they may have a cause of action against X. If X took steps to anonymize their identity, such as using Tor networks or encrypting their information,[154] Y may be unable to discover X's identity or even pursue the claim.[155] Even if X's identity is found, the cost of suing could be prohibitively expensive.[156] Finally, even if X faces charges or is found liable, Y will have limited means of preventing other versions of the video from spreading.

[43] Legal remedies do not surmount the three unique characteristics of malicious deepfake harm. First, individuals may see these deepfakes or believe they are true. Second, legal remedies seemingly do little to prevent the proliferation of deepfakes. By the time a plaintiff brings their case, overcomes the various legal hurdles listed above, and potentially receives a remedy, a deepfake will have spread across multiple platforms. Third, legal remedies cannot address the changing nature of deepfakes. As deepfake technologies continue to evolve, a flurry of lawsuits will only create inconsistent outcomes and caselaw.

---

[152] *See generally* Mary Anne Franks, *Drafting an Effective "Revenge Porn" Law: A Guide for Legislators*, CYBER CIVIL RIGHTS INITIATIVE (Sept. 22, 2016), https://www.cybercivilrights.org/guide-to-legislation/ [https://perma.cc/GAH8-C74P] (discussing the issues with removing nonconsensual content on the Internet given the failure of current laws, the difficulties of removal, and that "malicious individuals do not fear the consequences of their actions.").

[153] *See* Chesney & Citron, *supra* note 2, at 1792.

[154] *See id.*

[155] *Id.*

[156] *Id.*

IV. PROPOSALS SHOULD TARGET ONLINE PLATFORMS BUT NOT ELIMINATE SECTION 230

[44] Other scholars and policymakers have taken a different approach to mitigating deepfake harm. Rather than criminalize the creation of deepfakes, these advocates argue that online platforms should assume liability. Targeting online platforms is a good start to crafting effective policy, but many of these proposals seek to eliminate Section 230 protections, which would have dangerous consequences.

A. Online Platforms Should Be Responsible for Combatting Deepfakes

[45] Online platforms are "online sites and services that 'host, organize, and circulate users' shared content or social interactions for them,' without producing much of that content, built on an infrastructure for processing data'… [and that] moderate the content and activity of users."'[157] Online platforms enable, promote, and rely on social connectivity. Common examples include Facebook, Twitter, Reddit, though there are many others. More importantly, these entities fall within the larger umbrella of "Internet Service Providers" (ISPs) that are granted important protections.

[46] Online platforms should be responsible for addressing deepfakes. First, the spread of deepfakes is inherently tied to online platform proliferation. Second, online platforms are best suited to handle these issues and have the most resources to address deepfake threats.[158] Platforms can monitor how deepfake technology is changing, in-real time. Additionally, deepfakes do not affect all online platforms equally. The spread of a deepfake video on Pornhub versus an innocuous group on Reddit implicates different concerns and responses. Allowing these companies to create their

[157] ROBYN CAPLAN, CONTENT OR CONTEXT MODERATION? 8 (2018) (citation omitted).

[158] Chesney & Citron, *supra* note 2, at 1804–08 (stating other candidates like government administrative agencies and their "potential roles appear quite limited" given the scope of their jurisdiction and interest).

own policies creates a larger array of responses that may be more effective to address deepfakes. The ecosystem of online platforms provides laboratories of experimentation and innovation to address these issues.

[47] Third, online platforms establish a consistent entity to be held liable. Although important protections for speech, such as Section 230 of the Communications Decency Act, may make legal liability impractical, these platforms can still be held liable in other ways. As this Article will explore in more detail in Part V, public perception, market harm, and consumer transparency create powerful incentives for online platforms to address deepfakes.

B. Many Proposals Seek to Eliminate Protections Granted Under Section 230

[48] Although these advocates may get the "whom" right, their solutions fail for targeting Section 230 of the Communications Decency Act. Passed in 1996, Section 230 grants online platforms immunity for hosting harmful content, with an exception for content that violates federal criminal law, the Electronic Communications Privacy Act, or intellectual property law.[159] Section 230 states that, "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider."[160] Companies that fall under Section 230's scope include Internet Service Providers (ISPs) and online services that publish third-party content, which includes major online platforms.[161]

[159] *See* Chesney & Citron, *supra* note 2, at 1795.

[160] Communications Decency Act, 47 U.S.C. § 230(c)(1)(1998), https://www.govinfo.gov/content/pkg/USCODE-2011-title47/pdf/USCODE-2011-title47-chap5-subchapII-partI-sec230.pdf [https://perma.cc/HQ7G-EA2T].

[161] *See Section 230 of the Communications Decency Act*, ELECTRONIC FRONTIER FOUNDATION, https://www.eff.org/issues/cda230 [https://perma.cc/F9WN-5XZW].

[49] Advocates of Section 230 believe that its broad protections have allowed the internet to flourish.[162] Section 230 created the frameworks for sites like YouTube and Vimeo to host videos, Amazon and Yelp to offer user reviews, Craigslist to host ads, and social networking to exist generally.[163] Section 230 has acted as an important shield against parties seeking to hold platforms liable for harms created on their sites.[164]

[50] Critics of that shielding function have argued that Section 230 should be amended to impose liability on online platforms. The crux of these arguments is that these companies have not earned their immunity shield. Some argue that Section 230 should be eviscerated.[165] Others,

[162] *See id.*

[163] *See id.*

[164] *See* Jay M. Zitter, *Liability of Internet Service Provider for Internet or E-Mail Defamation*, 84 A.L.R. 5TH 169 (2000), https://1.next.westlaw.com/Document/Idf15c8c0469111daaeefbddf49df57ea/View/FullText.html?transitionType=Default&contextData=(sc.Default)&VR=3.0&RS=cblt1.0&__lrTS=20200423181227359&firstPage=true&bhcp=1&CobaltRefresh=3254 [https://perma.cc/H7DP-JWD7] (listing key issues and cases regarding Section 230); *see also CDA 230: Key Legal Cases*, ELECTRONIC FRONTIER FOUNDATION, https://www.eff.org/issues/cda230/legal [https://perma.cc/3P47-WBPK] (providing an overview of the key cases handling Section 230).

[165] *See Time to Reform CDA 230, Testimony to the House Subcommittee on Communication and Technology and the Subcommittee on Consumer Protection and Commerce*, 116th Cong., 4 (written testimony of Gretchen Peters, Executive Director of the Alliance to Counter Crime Online), https://energycommerce.house.gov/sites/democrats.energycommerce.house.gov/files/documents/Testimony_Peters.pdf [https://perma.cc/GM6G-HHUW] (stating that CDA 230 reforms should strip platforms' immunity "for hosting terror and serious crime content," increase the "onus on tech firms to monitor their platforms," "regulating that firms must report crime and terror activity, along with full data about users who uploaded it, to law enforcement" and other suggestions); *see also* Elliot Harmon, *Sen. Hawley's "Bias" Bill Would Let the Government Decide Who Speaks*, ELECTRONIC FRONTIER FOUNDATION (June 20, 2019), https://www.eff.org/deeplinks/2019/06/sen-hawleys-bias-bill-would-let-government-decide-who-speaks [https://perma.cc/PF59-ZT2A] (discussing a bill proposed by Sen. Josh Hawley that would strip certain platforms of their Section 230 immunity unless they

including Citron and Chesney, have argued for a moderate approach to the issue.[166] This Article discusses Citron and Chesney's popular approach[167] to suggest that the weaknesses of a moderate approach demonstrate that more extreme approaches will be futile.

[51] Citron and Chesney argue that a "reasonable steps" standard should be applied to online platforms.[168] They propose that, Section 230(c)(1) protections, or those that specifically state ISPs shall not be treated as publishers or speakers of information posted by their users, should be

prove they do not engage in political viewpoint discrimination); *see also Donald Trump – Twitter*, Factba.se.com (Jan. 29, 2021), https://factba.se/biden/topic/twitter?q=&f= [https://perma.cc/5MSB-LYN6] (demonstrating that President Trump has become an advocate of repealing Section 230, tweeting a confusing Executive Order regarding limiting Section 230 protections"); *see also* Exec. Order No. 13925, Preventing Online Censorship, 85 Fed. Reg. 106 34079 (May 28, 2020); *see also* U.S. DEP'T OF JUST., *Department of Justice's Review of Section 230 of the Communications Decency Act of 1996*, https://www.justice.gov/ag/department-justice-s-review-section-230-communications-decency-act-1996?utm_medium=email&utm_source=govdelivery [https://perma.cc/YRB4-U57W] (explaining that the Justice Department recently released proposed changes of Section 230 amendments that would reduce platform immunity).

[166] *See* Chesney & Citron, *supra* note 2, at 1799.

[167] *See Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, SSRN (July 14, 2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954 [https://perma.cc/7H3E-2MZ9] (explaining that Citron and Chesney's article is one of the leading papers on deepfakes, having been downloaded nearly 16,000 times, ranked 216 on SSRN, and already cited 37 times within a year of its publication); *see also Disinformation on Steroids: The Treat of Deep Fakes*, COUNCIL ON FOREIGN RELATIONS (Oct. 16, 2018), https://www.cfr.org/report/deep-fake-disinformation-steroids [https://perma.cc/LFW6-RGKW] (showing that the authors have written extensively on the issue, for a variety of platforms); *see also Deepfakes and the New Disinformation War*, FOREIGN AFFAIRS (Feb. 2019), https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war [https://perma.cc/Z6RA-YDKH].

[168] Citron & Chesney, *supra* note 2, at 1799.

"conditional rather than automatic."[169] In order for this protection to apply, "an entity must demonstrate that it has taken 'reasonable steps' to ensure that its platform is not being used for illegal ends."[170]

[52] Citron and Chesney acknowledge that it may be impossible for large ISPs and social platforms to immediately respond to complaints.[171] Their amendment would "minimize the most serious [deepfake] harms," and if the "reasonably available technical and other means for detection of harmful fakes are limited [,] so too will be the [platform's] obligation . . . ."[172] Thus, a reasonable steps standard would incentivize platforms to use improving technology.[173] However, Citron and Chesney's proposal ignores economic, cultural, and technological realities. Specifically, Citron and Chesney's reasonable steps standard is unrealistic for three reasons: (1) it imposes overly stringent duties on platforms, (2) it creates platform consolidation that would deter innovation, and (3) it would chill online speech.

1. A Reasonable Steps Standard Ignores the Complexity of Content Moderation

[53] Citron and Chesney's argument implies that online platforms have a duty to recognize and remove harmful fake content. That duty is split into two distinct, but interrelated, content moderation duties, each extraordinarily complex: (1) the duty to take reasonable steps to recognize fake content, and (2) the duty to take reasonable steps to remove this content.

[169] *Id.*

[170] *Id.*

[171] *Id.*

[172] *Id.* at 1799–1800.

[173] *See* Citron & Chesney, *supra* note 2, at 1800.

[54] Taking reasonable steps to recognize fake content is extraordinarily difficult. Citron and Chesney acknowledge that the accuracy rate of deepfake detection technology is not satisfactory.[174] More likely, technology and deepfakes will play a game of cat-and-mouse, evolution and reaction, a "back and forth" that "[i]ronically . . . . mimics the technology at the heart of the deepfakes: the generative adversarial network . . . ."[175] Depending on how platforms define deepfakes,[176] certain videos like the Nancy Pelosi "cheapfake" may not even qualify as a deepfake for removal. In a constantly evolving space, legislation that imposed a "reasonable step" – be it a specific moderation definition, process, code, or timeframe -would sow further confusion.[177]

[174] *Id.* at 1800 n.226 (stating that "[w]ith current technologies, it is difficult, if not impossible to automate the detection of certain illegal activity. That is certainly true of deep fakes in this current technological environment.").

[175] James Vincent, *Deepfake Detection Algorithms Will Never Be Enough*, THE VERGE, (June 27, 2019), https://www.theverge.com/2019/6/27/18715235/deepfake-detection-ai-algorithms-accuracy-will-they-ever-work [https://perma.cc/36V8-R8WR].

[176] *Id.* (explaining that under Facebook's Policy a cheapfake may not qualify for removal given it was created using machine learning or AI); *see also* Donie O' Sullivan, *Doctored Videos Shared to Make Pelosi Sound Drunk Viewed Millions of Times on Social Media*, CNN (May 24, 2019), https://www.cnn.com/2019/05/23/politics/doctored-video-pelosi/index.html [https://perma.cc/M8LH-BWG7] (describing the Nancy Pelosi "cheapfake" that took a video of the politician and slowed it down "by almost 75%" to make her appear drunk); The New York Times, *Deepfakes: Is This Video Even Real | NYT Opinion*, YOUTUBE (Aug. 14, 2019), https://www.youtube.com/watch?v=1OqFY_2JE1c [https://perma.cc/PK8S-WW6G] (a video of a technological expert likely using machine learning and AI programs to create a highly realistic impersonation of pop singer Adele); Britt Paris & Joan Donovan, *supra* note 121.

[177] Karni Chagal-Feferkorn, *The Reasonable Algorithm*, 2018 U. ILL. J.L. TECH. & POL'Y 111 (describing the complexities of developing a reasonableness standard for algorithms, including issues of compensation, deterrence, and defining how and to whom the standard applies); Ryan Abbott, *The Reasonable Computer: Disrupting the Paradigm of Tort Liability*, 86 GEO. WASH. L. REV. 1 (2018).

[55] A reasonable steps standard ignores the variety of content moderation approaches platforms use. A platform's content moderation approach depends on the platform's size, features, and business model.[178] The moderation needs of a search engine like Bing is different than the moderation needs of YouTube.[179] Even within similar platforms, approaches may vary among differing business and user models. Vimeo, a video-hosting platform, has largely been "spared disinformation" and "'fake news'" compared to peer sites like YouTube largely because Vimeo uses a subscription-based model and has many more professional users.[180]

[56] During the 2018 Content Moderation Conference in Santa Clara, CA, this variety of approaches became apparent as companies provided "Under the Hood" looks at their content moderation policies.[181] Pinterest revealed they had approximately "11.5 employees" responsible for moderating "200M+ MAUs, 100 Billion + Pins, and 30 + Languages."[182] In contrast, Google employed thousands of employees to "ensure compliance both with local laws and with Google's content policies."[183] These

[178] CAPLAN, *supra* note 157, at 9.

[179] *Id* at 10.

[180] *See id.*

[181] *See generally Content Moderation & Removal at Scale: Overview of Each Company's Operations*, SANTA CLARA UNIV. SCH. OF L. (Feb. 2, 2018, 10:00 AM), https://santaclarauniversity.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=6e2bf22d-52cd-4e3f-9324-a8810187bad7 [https://perma.cc/7VUW-ECGG] (summarizing how different Internet companies operationalize the moderation and removal of third-party generated content).

[182] Adelin Cai, *Pinterest Content Moderation Conference Slides*, SANTA CLARA UNIV. SCH. OF L. (Feb. 2, 2018, 10:00 AM), http://1x937u16qcra1vnejt2hj4jl-wpengine.netdna-ssl.com/wp-content/uploads/Cai-Content-Moderation-Slides.pdf [https://perma.cc/4PB6-PPZT].

[183] JEFF KOSSEFF, THE TWENTY-SIX WORDS THAT CREATED THE INTERNET 221 (2019); *See* April Glaser, *Want a Terrible Job? Facebook and Google May Be Hiring*, SLATE (Jan. 18, 2018, 11:44 AM), https://slate.com/technology/2018/01/facebook-and-google-

employees have a wide variety of backgrounds, work collaboratively, and can escalate difficult cases to lawyers, specialists, and experts.[184]

[57] A reasonable step standard ignores the realities that separate these approaches. For comparably sized companies with distinctly different products, a reasonable steps standard creates confusion: what is "reasonable" in terms of Facebook's content moderation policies may be different than Twitter's.[185] As the Vimeo and YouTube distinction shows, platforms that host similar types of content vary in scale and user base. It is unclear if "reasonable steps" should or could account for any given site's proclivity for hosting malicious deepfake content. Are the "reasonable steps" for Mind Geek, the major conglomerate that owns multiple pornographic sites, different than Vimeo's? Citron and Chesney state that "[t]he scope of the duty would need to track salient differences

are-building-an-army-of-content-moderators-for-2018.html [https://perma.cc/77CD-98GQ] (highlighting that Internet companies, including Google, are hiring thousands of employees to help moderate).

[184] *Id.*; *see also* Internet Society North American Bureau, *COMO SUMMIT 4 - Under the Hood: UGC Moderation (Part 2)*, YouTube (May 15, 2018), https://www.youtube.com/watch?v=SRR-xJAp0j0&list=PL4buVHalBRoMgSatKZoIj0vy4LjNP-iaz&index=4 [https://perma.cc/6873-6THZ] (the panel of experts from Google, GitHub, the Wikimedia Foundation, and Facebook describing their content moderation policies at the Content Moderation at Scale Summit in Washington D.C.).

[185] *See* Jane C. Hu, *Twitter Has Set Itself Up for an Enormous New Content Moderation Problem*, SLATE (Nov. 20, 2020), https://slate.com/technology/2020/11/twitter-fleets-content-moderation-misinformation-harassment.html [https://perma.cc/C32Q-NFHC] (describing how Twitter's new "Fleet" function that allowed users to upload new video content may create moderation loopholes with banned imagery); Paul M. Barrett, *Who Moderates the Social Media Giants?*, NYU STERN CENTER FOR BUSINESS AND HUMAN RIGHTS (June 2020), https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/5ed9854bf618c710cb55be98/1591313740497/NYU+Content+Moderation+Report_June+8+2020.pdf [https://perma.cc/WMW7-PCMW] (stating that outsourcing content moderation issues has led to human rights issues).

among online entities."[186] Unfortunately, this tracking would likely come too little too late, be imprecise, and have a disastrous effect on platforms.

[58] The second duty to take reasonable steps to remove content is an even harder challenge. Assuming an online platform has identified malicious deepfake content, they must also reasonably determine whether to *remove* the content.

[59] Consider the following hypothetical spread of a deepfake video across multiple platforms: a deepfake is posted on Instagram by known deepfake creator, "@the_fakening." The deepfake uses AI technology to portray Bernie Sanders smoking marijuana, criticizing President Trump and the Democratic establishment with profanity. @Daquan, a popular "meme" account on Instagram with 14.7 million followers, reposts the content with a short caption saying, "Bernie going in" and attributing the post to @the_fakening. For many of @Daquan's followers, this post would likely be recognized as satire and manipulated media, given the page's reputation for posting satirical content, memes, and the attribution to @the_fakening.

[60] However, the video begins to spread like wildfire. Several other aggregator accounts on Instagram begin to repost the video, altering the content to include new commentary and text such as, "Bernie Sanders unhinged," "Bernie Sanders unleashed," and "Shocking New Video of Bernie." On Reddit, it trends in the "r/politics" subgroup, with many users recognizing the work is manipulated but upvoting it for its humor and commentary. On Facebook, far-right groups post the video, asserting its truthfulness. Democrat groups denounce the video as propaganda meant to inflame voters. Other groups reshare the video adding additional commentary about how it is an exemplary deepfake, their own thoughts on politics, and more. Bernie Sanders releases his own statement denouncing the videos as false. Eventually, Donald Trump, Jr. reposts the video on his Twitter stating "Wow! Crazy Bernie Sanders Really Is Crazy! #TRUMP2024."

[186] Chesney & Citron, *supra* note 2, at 1799.

[61] Assume that platforms have the technology to detect these deepfakes immediately and accurately. At what point would it become a "reasonable step" to remove these videos? Would Instagram have to delete the video when it was posted by @the_fakening and reposted on @Daquan? Likely not, as the post is clearly political commentary or a parody. Finding it reasonable to remove the video at this point would be overinclusive and take an approach to deepfakes that implicates too much innocuous content.

[62] When and how would intervention be reasonable? Would it be a "reasonable step" for platforms to remove a viral video as it begins spreading? Would it be a "reasonable step" to remove the video unilaterally on all platforms when only certain platforms' users assert the veracity of the video? Would it be a "reasonable step" to immediately demarcate the video as manipulated when it is uploaded? These questions demonstrate just a few of the difficulties of imposing a reasonable steps standard on content moderation duties.

2. A Reasonable Steps Standard Would Lead to Consolidation and Confusion

[63] Citron and Chesney do not articulate how courts or enforcement agencies would distinguish between different platforms, but state there should be a "separate rule for websites designed to facilitate illegality in contrast to large ISPs linking millions to the Internet."[187] A legal rule flexible enough to account for the diversity and variety of online platforms is unlikely. A reasonable steps standard alone would likely further platform consolidation, favoring larger companies, and create confusing standards in enforcement.

[64] Even if courts granted cohesive rulings, consolidation could occur if online platforms were held to a reasonable-steps standard. Imagine a plaintiff sues Facebook for the spread of a malicious deepfake. Facebook's conduct is found unreasonable because Facebook did not utilize a specific technology that could have detected the video and removed it faster. The

[187] *Id.* at 1800.

court issues an injunction requiring Facebook to remove the video and obtain better technology. Depending on the court's language, a broad pronouncement of what is "reasonable" could force hundreds of companies to reassess their strategies overnight. Even a limited ruling could force smaller competitors, that do not have the appropriate resources to invest in newer technologies, out of business.

[65] Another potential scenario is that Facebook's approach *is* found reasonable. That ruling also sends a green light to other large platforms to quickly adopt these practices. But depending on the court's ruling, many smaller platforms could be left in disarray. Given their size, audience, and revenue, would they be forced to comply? Employing hundreds of people and multiple teams may be feasible for platforms like Facebook and Google, but it is likely impossible for other popular platforms with teams of ten people or less like Medium, Discord, or Patreon.[188]

[66] Larger companies may use opportunities like the hypothetical Facebook ruling to consolidate with smaller competitors or offer their resources in ways that force smaller platforms to collaborate. These dynamics may lead larger platforms to require smaller platforms to pay them high fees, narrow their audiences, or agree to terms that otherwise diminish the small platforms.[189]

---

[188] CAPLAN, *supra* note 157, at 11.

[189] *See, e.g.*, George P. Slefo, *Ad-Tech Consolidation Posted to Accelerate Under GDPR*, ADAGE (Dec. 6, 2017, 9:00 AM), https://adage.com/article/digital/ad-tech-consolidation-poised-acceleration-gdpr/311535 [https://perma.cc/E8AV-E42Z] (stating that the increased burdens the GDPR places on companies will create "no room for exchanges that simply sit in the middle and add no value" in ad tech); Nicholas Martin, et. al., *How Data Protection Regulation Affects Startup Innovation*, INFORMATION SYSTEMS FRONTIERS 21, 1307, 1319, https://link.springer.com/content/pdf/10.1007/s10796-019-09974-2.pdf [https://perma.cc/QXB5-T9L5] (stating that data protections, like the GDPR, can discourage entrepreneurship and lead to product abandonment); Eline Chivot & Daniel Castro, *What the Evidence Shows About the Impact of the GDPR After One Year*, CENTER FOR DATA INNOVATION S (June 17, 2019), https://www.datainnovation.org/2019/06/what-the-evidence-shows-about-the-impact-of-the-gdpr-after-one-year [https://perma.cc/J7U3-UDX4] (stating that data protections, like the GDPR, can

[67] Ultimately, consolidation and court findings of reasonableness would effectively undermine the innovation and flexibility demanded by Section 230 that is needed to combat malicious deepfakes. Finding specific avenues and methods of addressing deepfakes "unreasonable" may cut off research on valuable opportunities for detection. Additionally, finding that a platform is responsible and should have taken a specific step may create confusion, even incentivizing platforms to be less responsive to user concern and harm.[190] This deterrence of innovation from platforms is the *exact* behavior that Section 230 sought to avoid. As Ron Wyden, one of the authors of Section 230 stated: "Chipping away . . . [at Section 230] will curtail the culture of innovation and bare-knuckled competition that have

discourage entrepreneurship and lead to product abandonment); Alec Stapp, *GDPR After One Year: Costs and Unintended Consequences*, TRUTH ON THE MARKET (May 24, 2019), https://truthonthemarket.com/2019/05/24/gdpr-after-one-year-costs-and-unintended-consequences [https://perma.cc/9UWD-6RYZ] (listing various effects of the GDPR including market consolidation, forgone investments and research, and multiple small and medium-sized businesses leaving the EU).

[190] KOSSEFF, *supra* note 183, at 166–205 (explaining that the line of reasoning of what a platform's duties are, and what content the platform is responsible for, is a question that courts now "continue to struggle with" and that judicial decisions in the past two decades have found platforms liable for failing to take specific steps if they were found to be an actual "information content provider" or acted by "publishing or speaking," thus losing their Section 230 immunity); *see also, e.g.*, *Fair Housing Council of Fernando Valley v. Roomates.Com, LLC*, 521 F.3d 1157, 1171–76 (9th Cir. 2008) (en banc) (holding that Roomates.com may be liable for asking illegal questions and publishing responses that may violate antidiscrimination housing laws). *Cf. See Barnes v. Yahoo!, Inc.*, 570 F.3d 1096, 1106-09 (9th Cir. 2009) (holding that a plaintiff might sue Yahoo under a theory of promissory estoppel given their lack of removing a profile continuing her nude pictures, despite Yahoo's Director of Communications confirming her request). By "allowing plaintiffs… to sue online platforms by carving out exceptions to Section 230, courts often discourage platforms from taking affirmative steps to prevent offensive online content. Had Yahoo simply ignored . . . [the plaintiff's] repeated request for help in removing the profiles . . . the Ninth Circuit probably would have affirmed the dismissal of the entire lawsuit . . . . [Thus] any outreach to customers are actions that go beyond mere editing and could place the platforms outside Section 230's protections." KOSSEFF, supra 183, at 195; *F.T.C. v. Accusearch Inc.*, 570 F.3d 1187, 1201 (10th Cir. 2009) (finding the defendant platform an internet content provider, thus outside the scope of Section 230, through its "actions [that] were intended to generate" offensive content).

been the defining characteristics of the internet for more than two decades."[191]

3. A Reasonable Steps Standard Would Chill Speech

[68] Citron and Chesney greatly underestimate the potential effects a reasonable steps standard would have on speech.[192] One often-predicted effect of Section 230 reform is that it would lead ISPs and online platforms to overregulate to escape liability.[193] This concern stems directly from the idea that it may be easier and cheaper for online platforms to take down all content that could potentially induce liability, rather than be selective.

[69] The few exceptions to Section 230 have shown that these concerns are warranted. In 2018, President Trump signed into law FOSTA-SESTA, or The Allow States and Victims to Fight Online Sex Trafficking Act of 2017, and the Stop Enabling Sex Traffickers Act.[194] FOSTA-SESTA was passed with bipartisan support to prevent sex trafficking by creating a Section 230 exception that makes website publishers responsible for third parties posting ads for prostitution, even consensual sex work, on their

[191] Ron Wyden, *Floor Remarks: CDA 230 and SESTA,* MEDIUM (Mar. 21, 2018), https://medium.com/@RonWyden/floor-remarks-cda-230-and-sesta-32355d669a6e [https://perma.cc/6UP8-ES65].

[192] *See* Chesney & Citron, *supra* note 2, at 1800 (acknowledging the effects on speech only briefly in their article, stating that their proposal "might drive sites to shutter (or to never emerge), and it might cause undue private censorship . . . .").

[193] *See, e.g.*, Hayley Tsukayama, et. al., *Congress Should Not Rush to Regulate Deepfakes*, ELECTRONIC FRONTIER FOUNDATION (June 24, 2019), https://www.eff.org/deeplinks/2019/06/congress-should-not-rush-regulate-deepfakes [https://perma.cc/HP4V-BKFP] (stating that [a]ltering Section 230's language to increase liability for harmful deepfakes will . . . sweep up contributions to the public discourse, like parodies and satires, but it will also implicate a range of other forms of lawful and socially beneficial speech, [given the incentives for] platforms [to] censor more.").

[194] Allow States and Victims to Fight Online Sex Trafficking Act of 2017, 47 U.S.C. § 230, 115th Cong (April 11, 2018), https://www.congress.gov/115/plaws/publ164/PLAW-115publ164.pdf [https://perma.cc/U2C7-453P].

platforms.[195] Although the bill targets an important cause, it was widely criticized for its effect on speech and sex workers.[196] Multiple platforms fundamentally restructured to limit speech in fear that just *knowing* of content would expose them to liability. For example, Craigslist shut down its "Personals" sections,[197] Tumblr banned "adult content,"[198] and Facebook banned "implicit sexual solicitation."[199] These reactions hurt the very communities the law strove to protect. Sex workers, anti-trafficking

[195] Aja Romano, *A New Law Intended to Curb Sex Trafficking Threatens the Future of Internet as we Know it*, VOX (Jul. 2, 2018, 1:08 PM), https://www.vox.com/culture/2018/4/13/17172762/fosta-sesta-backpage-230-internet-freedom [https://perma.cc/QV4U-Z4B4]. *See* Allow States and Victims to Fight Online Sex Trafficking Act of 2017, at §2(1), §242(1)(A)(a—b) (explicitly stating: "[S]ection 230 of the Communications Act…was never intended to provide legal protection to websites that unlawfully promote and facilitate prostitution . . . [,]traffickers . . . and have done nothing to prevent…victims of force, fraud, and coercion . . . ."); *Id.* at § 2(1–2).

[196] Jaimee Bell, *FOSTA – SESTA: Have controversial sex trafficking acts done more harm than good?*, BIG THINK (Jan. 22, 2021), https://bigthink.com/politics-current-affairs/fosta-sesta-sex-trafficking?rebelltitem=1#rebelltitem1 [https://perma.cc/3GNA-NWVD].

[197] *See* Aja Romano, *A New Law Intended to Curb Sex Trafficking Threatens the Future of Internet as we Know it*, VOX (Jul. 2, 2018, 1:08 PM), https://www.vox.com/culture/2018/4/13/17172762/fosta-sesta-backpage-230-internet-freedom [https://perma.cc/QV4U-Z4B4].

[198] Alexander Cheves, *The Dangerous Trend of LGBTQ Censorship on the Internet*, OUT MAGAZINE (Dec. 6, 2018), https://www.out.com/out-exclusives/2018/12/06/dangerous-trend-lgbtq-censorship-internet [https://perma.cc/WK9P-NDFK].

[199] Elliot Harmon, *Facebook's Sexual Solicitation Policy is a Honeypot for Trolls*, ELECTRONIC FRONTIER FOUNDATION (Dec. 7, 2018), https://www.eff.org/deeplinks/2018/12/facebooks-sexual-solicitation-policy-honeypot-trolls [https://perma.cc/72W2-T3F2].

organizations,[200] and the Department of Justice[201] largely opposed the bill. Critics argued the bill conflated issues of nonconsensual and consensual sex work and forced sex workers to pursue unsafe work.[202] These critiques led Senator Elizabeth Warren to propose an act examining the efficacy of FOSTA-SESTA given the "significant impacts on the health and safety of people who engage in consensual, transactional sex."[203] Danielle Citron, a major proponent of amending Section 230, acknowledged that "FOSTA endorses a piecemeal approach to a problem that should be solved more comprehensively."[204]

[200] *See, e.g.*, *Freedom Network Urges Caution in Reforming the CDA*, FREEDOM NETWORK (Sept. 18, 2017), https://www.eff.org/files/2017/09/18/sestahearing-freedomnetwork.pdf [https://perma.cc/X73R-GAC7] (stating that "amending Section 230" could "deter responsible website administrators from trying to identify and report trafficking"); Alex Andrews, *SWOP-USA Stands in Opposition of Disguised Internet Censorship Bill SESA, S.* 1963, SWOP USA (Aug. 11, 2017), https://web.archive.org/web/20171024095814/http://www.new.swopusa.org/2017/08/11/call-to-actionpress-release-swop-usa-stands-in-direct-opposition-of-disguised-internet-censorship-bill-sesta-s-1963-call-your-state-representatives-and-tell-them-to-fight/ [https://perma.cc/4428-3GQ3] (suggesting that SESTA would compromise the fight against sex trafficking and harm sex workers).

[201] Letter from Stephen E. Boyd, Assistant Att'y Gen., Dep't. Justice, to Hon. Robert Goodlatte, Chairman, Comm. on the Judiciary (Feb. 27, 2018), https://assets.documentcloud.org/documents/4390361/Views-Ltr-Re-H-R-1865-Allow-States-and-Victims.pdf [https://perma.cc/SJC5-4TNK].

[202] *See, e.g.*, Ana Valens, *SESTA-FOSTA is 'detrimental' to sex workers' safety, study confirms*, DAILY DOT (JAN 27, 2021), https://www.dailydot.com/irl/sesta-fosta-report-sex-work/ [https://perma.cc/QY8E-6ERN].

[203] SESTA/FOSTA Examination of Secondary Effects for Sex Workers Study Act, S. 3165, 116th Cong. §2(10) (2020).

[204] Danielle Citron & Quinta Jurecic, *FOSTA: The New Anti-Sex Trafficking Legislation May Not End the Internet, But It's Not Good Law Either*, LAWFARE (Mar. 28, 2018), https://www.lawfareblog.com/fosta-new-anti-sex-trafficking-legislation-may-not-end-internet-its-not-good-law-either [https://perma.cc/A2BR-4YR6].

[70] Regardless of one's views on sex work, an important lesson can be gleamed from FOSTA-SESTA: amending Section 230 has unintended consequences that can powerfully curb speech and hurt those that the law seeks to protect. The potential for overinclusive, negative effects is not a probability but a certainty. Imposing a reasonable steps standard could lead platforms to ban many forms of speech that are needed to address the harms of deepfakes. For example, a team of artists created a deepfake of Mark Zuckerberg to criticize Facebook's policy decisions toward deepfakes.[205] The video has racked up hundreds of thousands (if not millions) of views, and led to public discussions and awareness of Facebook's decisions, and deepfakes in general.[206] A policy that bans deepfakes too broadly, as is probable under a reasonable steps standard, would curb these important uses of deepfakes and other manipulated media as forms of social commentary, critique, and discourse.

[71] Ultimately, giving companies the responsibility of removing content grants these platforms the authority to control our speech. Online platforms already have immense control over the way we handle our content and what we see as "true."[207] This has already led to accusations of political

[205] Samantha Cole, *This Deepfake of Mark Zuckerberg Tests Facebook's Fake Video Policies*, MOTHERBOARD: TECH BY VICE (June 11, 2019), https://www.vice.com/en_us/article/ywyxex/deepfake-of-mark-zuckerberg-facebook-fake-video-policy [https://perma.cc/935K-A8WD].

[206] *See, e.g.*, *id.*; Multimedia LIVE, *Artists create Zuckerberg 'deepfake' video*, YOUTUBE (Jun. 13, 2019), https://www.youtube.com/watch?v=cnUd0TpuoXI [https://perma.cc/6BT6-JCKF]; Makena Kelly, *Instagram will leave up deepfake video of Mark Zuckerberg*, THE VERGE (Jun. 11, 2019), https://www.theverge.com/2019/6/11/18662027/instagram-facebook-deepfake-nancy-pelosi-mark-zuckerberg [https://perma.cc/Y5UU-5QUE].

[207] *See* Kalev Leeatru, *Social Media Platforms Will Increasingly Define 'Truth'*, FORBES (Aug. 24, 2019, 6:32 PM), https://www.forbes.com/sites/kalevleetaru/2019/08/24/social-media-platforms-will-increasingly-define-truth/?sh=6fbc58fd6427#67683a [https://perma.cc/UJ9M-EHM5] ("[Social media companies] are increasingly claiming the right to define truth itself . . . [and] control the societal discourse of the modern age.").

censorship,[208] partisanship,[209] and silencing of viewpoints.[210] A fixed "reasonable" standard of what is malicious in this developing space coupled with enforcement by platforms will grant platforms unbridled power. Online platforms will become the creators of truth, a responsibility that we should be wary to grant.

4. What Can We Learn From the "Reasonable Steps" Standard?

[72] It is important to note the strengths of Citron and Chesney's proposal. Both authors recognize that Section 230 reform is not the silver bullet to end deepfake harm. They state that "features used to control the scope of platform liability are only a partial solution to the deep-fakes challenge. Other policy responses will be necessary."[211] Indeed, other policy responses that prioritize extensive collaboration and platform accountability are needed.

[73] The reasonable steps standard handles malicious deepfakes better than all the other aforementioned proposals. Online platforms would likely

[208] *See, e.g.*, Casey Newton, *It Turns Out There Really is an American Social Network Censoring Political Speech*, THE VERGE (Sept. 26, 2019, 6:00 AM), https://www.theverge.com/2019/9/26/20883993/tiktok-censorship-china-bytedance-politics [https://perma.cc/CG63-B2UG] (describing social media platform TikTok's alleged efforts to censor political speech).

[209] *See* Christopher A. Bail et al., *Exposure to Opposing Views on Social Media Can Increase Political Polarization*, 115 PROC. NAT'L ACAD. SCIS. 9216, 9216 (2018) (finding that exposure to social media "echo chambers" can further entrench partisanship when faced with opposing views).

[210] *See, e.g.*, Makena Kelly, *White House Launches Tool to Report Censorship on Facebook, YouTube, Instagram, and Twitter*, THE VERGE (May 15, 2019, 5:18 PM), https://www.theverge.com/2019/5/15/18626785/white-house-trump-censorsip-tool-twitter-instagram-facebook-conservative-bias-social-media [https://perma.cc/3F5Y-YB75] (discussing new initiatives by the White House to counter allegations of conservative censorship on social media platforms).

[211] Chesney & Citron, *supra* note 2, at 1800–01.

overly penalize and remove videos that are deemed manipulated. Users that flag violating videos would also likely see more of these videos taken down without question. Companies would invest more in technologies that broadly recognize and remove manipulated media. Less videos would be seen, so upfront harm would be reduced. If videos escape detection, flagging them would likely lead to a prompt removal. This would mitigate the proliferation of deepfakes on social media. As a result of such a "scorched earth" policy, the everchanging definition and technologies of deepfakes may cease to be an issue. But at what cost?

[74] Although some deepfakes are malicious, one must respect the values of discourse, speech, and expression that are central to a democracy. Ultimately, the reasonable steps approach fails for many practical reasons but largely for the principle that it does not give sufficient weight to the need for innovation by platforms, flexibility of approach, or protections for speech that foster true solutions to deepfakes. As a result, a better solution must be found.

V. A NOVEL PROPOSAL: DISCLOSURES, COLLABORATION, AND EDUCATION

[75] A review of the arguments demonstrates that any viable solution must address that deepfake harm is (1) inherently upfront; (2) rapidly spreading on social media; and is (3) difficult to combat given the changing definition and technology behind deepfakes. Any proposal to combat deepfakes should also (4) address online platforms; but (5) not amend key internet protections, such as Section 230. This Article suggests a tripartite proposal to increase transparency, promote collaboration among the government and public sector, and create extensive public education resources about deepfakes and manipulated media.

A. Part I: Transparency Disclosures

[76] The first part of this proposal mandates that online platforms provide annual transparency disclosures to their users regarding deepfake and manipulated media. Many advocates and critics of online platforms have

noted the need for more transparency. The push for transparency may be self-serving. For "Section 230 to survive future challenges, platforms must not only improve their moderation practices but also publicly explain how they do so."[212] Platforms are also beginning to recognize the importance of transparency, as several companies are releasing more information to their users.[213]

[77] In 2018, stakeholders agreed on content moderation principles, the Santa Clara Principles, for online platforms.[214] These principles were "meant to serve as a starting point, outlining minimum levels of transparency and accountability" for online platforms.[215] They included creating more transparency for the "numbers of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines", providing "notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension", and "provid[ing] a meaningful opportunity for timely appeal of any content removal or account suspension."[216]

[78] Platforms should follow a set of transparency disclosures in line with the Santa Clara principles regarding their practices and takedowns of deepfakes and manipulated media. More specifically, platforms should include:

---

[212] KOSSEFF, *supra* note 183, at 250.

[213] Andrew Crocker et al., *Who Has Your Back? Censorship Edition 2019*, ELEC. FRONTIER FOUND. (June 12, 2019), https://www.eff.org/wp/who-has-your-back-2019#executive-summary [https://perma.cc/NPG9-EMRQ] (analyzing the transparency and content moderation policies of many major online platforms).

[214] THE SANTA CLARA PRINCIPLES (Jan. 31, 2021), https://www.santaclaraprinciples.org [https://perma.cc/C6WN-X9BF] (discussing principles created following the 2018 Content Moderation at Scale conference in Santa Clara, CA, and the second Content Moderation at Scale conference in Washington, D.C.).

[215] *Id.*

[216] *Id.*

- How they define deepfakes and manipulated media and, if these definitions were changed recently, how they were changed;
- an assessment of how deepfake content and manipulated media has been used on their websites including:
    - specific case-examples of malicious and non-malicious use;
    - parties that may be using and creating these videos;
- the platform's current policy toward manipulated media and, if these policies were recently changed, how they were changed including details on:
    - the platform's notice policies, including:
        - how the platform provides notice;
        - average response times to videos;
        - how these videos are being detected;
    - the platform's appeal policies, including:
        - how the company provides an appeal process, or if not, why;
        - the number of successful appeals;
        - the number of unsuccessful appeals;
        - the average time of an appeal;
- any changes in legislation or law that has affected the online platform;
- a thorough assessment of manipulated media content that was reported to the company including,
    - the total number of discrete posts and accounts flagged;
    - the total number of discrete posts removed and accounts suspended;
    - the number of discrete posts and accounts flagged and number of discrete posts removed and accounts suspended, by category of policy or rule violated;
    - the number of discrete posts and accounts flagged and number of discrete posts removed and accounts suspended, by category of legal request;

- the number of discrete posts and accounts flagged, and number of discrete posts removed, and accounts suspended, by format of content at issue (e.g. text, audio, image, video, live stream);
- the number of discrete posts and accounts flagged, and number of discrete posts removed, and accounts suspended, by source of flag (e.g. governments, trusted flaggers, users, automated detection);
- the number of discrete posts and accounts flagged, and number of discrete posts removed and accounts suspended, by locations of flaggers and impacted users (where apparent); and
- any other information as deemed relevant by online platform or legislature.[217]

[79] The report should be freely and publicly available to all users of the platform, on at least an annual basis. All users should receive notifications that the reports are available. Ideally, this means that users will be emailed the report and receive a notification when they log into the platform.[218]

[80] Failure to comply with these disclosures should lead to liability for online platforms, though liability should not lead to widescale removals of protections for speech, such as amending Section 230. Instead, liability should be based on similar, but less aggressive, structures that are used in

[217] *See id.* (suggesting metrics that are the "ideal" data that platforms could provide).

[218] Several large platforms already provide transparency results. *See, e.g.*, *Facebook Transparency Report*, FACEBOOK, https://transparency.facebook.com/ [https://perma.cc/W4B3-W87N]; *Transparency*, REDDIT, https://www.reddit.com/wiki/transparency [https://perma.cc/4S6V-DZQD]; *Google Transparency Report*, GOOGLE, https://transparencyreport.google.com/?hl=en [https://perma.cc/2KLM-6QK8]. However, many of these reports contain varying levels of details and information. *See* Andrew Crocker et al., *supra* note 213 (stating that Facebook's reports do not "report the total number of government takedown requests received" and provide a limited category of Community Standard takedown requests). Additionally, there is no indication that online platforms tell their users such reports have been completed as a notification.

the GDPR and CCPA (the California Consumer Privacy Act). GDPR fines can be expansive: upwards of 20 million euros, or even 4% of the entities "total global turnover of the preceding fiscal year."[219] These may vary according to the case, the nature of the company, and the nature of the violation in order to have punishments that are "effective, proportionate and act as a deterrent."[220] CCPA fines may be up to a maximum of $7500 for "each intentional violation," with businesses given the opportunity to cure any violation within 30 days.[221] Given the breadth of some platforms, this could reach millions of dollars for violations that affect multiple users.

[81] Transparency disclosures work well to mitigate deepfake harms. First, these disclosures require online platforms to take an affirmative step in researching manipulated media and deepfakes on their sites. Although many platforms seem to be taking tentative steps to research these issues, a transparency disclosure mandates platforms do so on an ongoing basis. Thus, platforms are required to evaluate their methods, policies, definitions, and efficiencies.

[82] Second, these disclosures directly combat the malicious profit motive of platforms. The profit motive states that online platforms are disincentivized from self-regulating and taking an active role in content moderation.[222] Various arguments posit that online platforms will fail to

---

[219] *GDPR Fines / Penalties*, INTERSOFT CONSULTING, https://gdpr-info.eu/issues/fines-penalties/ [https://perma.cc/24TA-UPW3].

[220] *Id.*

[221] California Consumer Privacy Act of 2018, S. 1121 at §1798.155(b), 2018 Cal. Sen. (2018), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1121 [https://perma.cc/9ECH-HYCC].

[222] *See* Peter Cohan, *Does Facebook Generate Over Half of Its Ad Revenue From Fake News?*, FORBES (Nov. 25, 2016, 10:36 AM), https://www.forbes.com/sites/petercohan/2016/11/25/does-facebook-generate-over-half-its-revenue-from-fake-news/?sh=1d6dd037375f [https://perma.cc/UY4K-MRX5] (describing how Facebook's fake news posts get more interaction than real news); *A Country in Crisis: How*

regulate malicious manipulated media because they are profitable.[223] More specifically, malicious deepfakes are the kind of inflammatory content that drive up media platform shares, ad revenues, and general usage.[224]

[83] Recent actions by companies have questioned the validity of the profit motive. Online platforms have shown a willingness to change policies due to their concerns that "further ignorance" of ineffective policies "would risk ruining the company's reputation with customers – and [their] source of revenues."[225] For example, Pinterest was concerned that it may be contributing to eating disorders, so it collaborated with the National Eating

Disinformation Online is Dividing the Nation: *Hearing Before the Subcomm. on Commc'n & Tech. and the Subcomm. on Consumer Prot. & Com.*, 116th Cong. (2020), https://docs.house.gov/meetings/IF/IF17/20200624/110832/HHRG-116-IF17-Wstate-FaridH-20200624.pdf [https://perma.cc/UQG7-3CBQ] (statement of Hany Farid, Professor, Univ. Cal.) (describing how platforms' "algorithmic amplification" recommends fake and dangerous content to engage users rather than truthful content).

[223] Anna Romero, *Shanmugam: Law Against Fake News Necessary Because Social Media Firms Put Profits First*, THE INDEPENDENT (Jan. 13, 2021), https://theindependent.sg/shanmugam-law-against-fake-news-necessary-because-social-media-firms-put-profits-first/ [https://perma.cc/26JN-3TMM].

[224] *See* Marie Boran, *How Social Media Platforms Battle Misinformation While Profiting From It*, IRISH TIMES (Feb. 6, 2020), https://www.irishtimes.com/business/technology/how-social-media-platforms-battle-misinformation-while-profiting-from-it-1.4160387 [https://perma.cc/4PQE-PKL4] (describing the difficulties of having platforms regulate and battle misinformation when they benefit these platforms by increasing viewers); Peter Cohan, *Does Facebook Generate Over Half of Its Ad Revenue From Fake News?*, FORBES (Nov. 25, 2016, 10:36 AM), https://www.forbes.com/sites/petercohan/2016/11/25/does-facebook-generate-over-half-its-revenue-from-fake-news/?sh=1d6dd037375f [https://perma.cc/UY4K-MRX5] (describing how Facebook's fake news posts get more interaction than real news) (discussing how Facebook likely makes revenue from fake news ads); Yaël Eisenstat, *I Worked On Political Ads at Facebook. They Profit By Manipulating Us*, WASH. POST (Nov. 4, 2019, 6:00 AM), https://www.washingtonpost.com/outlook/2019/11/04/i-worked-political-ads-facebook-they-profit-by-manipulating-us/ [https://perma.cc/4WS4-EVAF] (stating how Facebook profits partly by amplifying lies and selling dangerous targeting tools).

[225] KOSSEFF, *supra* note 183, at 241.

Disorder Association to compile " a list of keywords related to the problem" that limit the search results and inform content removal decisions.[226] Online platforms have made their own policy decisions specifically for deepfakes due to increased scrutiny. Facebook, for example, instituted a "deepfake ban"[227] amid much criticism[228] and announced the creation of a content "Oversight Board" to review content moderation decisions[229] in the wake of further backlash.

[84] Transparency disclosures directly play to the dynamic between online platforms and their users. By mandating platforms give consumers information, it institutes a consumer check on platforms.[230] This check can

[226] *Id.* at 243; *see also* Carolyn Gregoire, *Pinterest Removes Eating Disorder-Related Content, Pro-Anorexia Community Continues to Thrive*, HuffPost (Aug. 10, 2012), https://www.huffpost.com/entry/pinterest-removes-eating-disorder-content_n_1747279 [https://perma.cc/W8XD-9HW9].

[227] Monika Bickert, *Enforcing Against Manipulated Media*, FACEBOOK (Jan. 6, 2020), https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/ [https://perma.cc/TZ7X-QNGH].

[228] Makena Kelly, *Facebook's Deepfake Ban Isn't Winning Over Critics*, THE VERGE (Jan. 7, 2020), https://www.theverge.com/2020/1/7/21055283/facebook-deepfake-ban-political-ads-shallowfakes-rules-moderation [https://perma.cc/7RT5-ERAZ] (describing how individuals have stated Facebook's ban does not go far enough and was confusing to many); James Vincent, *Facebook's Problems Moderating deepfakes Will Only Get Worse in 2020*, THE VERGE (Jan. 15, 2020), https://www.theverge.com/2020/1/15/21067220/deepfake-moderation-apps-tools-2020-facebook-reddit-social-media [https://perma.cc/ZQ38-6ACW] (criticizing Facebook, and other platforms, policies toward deepfakes).

[229] Brent Harris, *Establishing Structure and Governance for an Independent Oversight Board*, FACEBOOK (Sep. 17, 2019), https://about.fb.com/news/2019/09/oversight-board-structure/ [https://perma.cc/CAC6-RGPS].

[230] *See* Mark MacCarthy, *How Online Platform Transparency Can Improve Content Moderation and Algorithmic Performance*, BROOKINGS (Feb. 17, 2021), https://www.brookings.edu/blog/techtank/2021/02/17/how-online-platform-transparency-can-improve-content-moderation-and-algorithmic-performance/ [https://perma.cc/43ZS-PK58].

have powerful repercussions that spur action faster, and more flexibly, than legal liability. One of the most expensive GDPR fines was $62,814,221 against Google for lack of transparency and valid consent in their practices.[231] In contrast, Facebook lost nearly $109 billion from its market valuation due to the Cambridge Analytica scandal, which exposed the company's controversial consumer data sharing practices.[232] This scandal fundamentally changed the way the company was perceived by the public and how it interacted with consumers. Consumers became much more critical of Facebook's data practices.[233] Although daily and monthly active users have increased, a statistic that may mean users have still retained their accounts and logged on, there was a nearly 20% decline of action, such as likes, shares, and posts, on Facebook.[234] Additionally, Facebook built new privacy protections for consumers, including making privacy and data

[231] Simon Fogg, *€50 Million Google GDPR Fine - Losers and Lessons from the GDPR*, TERMLY (May 17, 2019), https://termly.io/resources/articles/google-gdpr-fine/ [https://perma.cc/E3KP-BSCV].

[232] Vaughn Highfield, *Facebook Suffers €83 Billion Financial Blow Following Cambridge Analytica Scandal*, ALPHR (Jul. 26, 2019), https://www.alphr.com/facebook/1009757/facebook-suffers-83-billion-financial-blow-cambridge-analytica-scandal [https://perma.cc/ZX29-257A].

[233] *See* Julia Carrie Wong, *The Cambridge Analytica Scandal Changed the World - But it Didn't Change Facebook*, THE GUARDIAN (MAR. 18, 2019), https://www.theguardian.com/technology/2019/mar/17/the-cambridge-analytica-scandal-changed-the-world-but-it-didnt-change-facebook [https://perma.cc/H2LT-HHBX]; *see also* Tom Gerken, *Whatsapp Co-founder Says it is Time to Delete Facebook*, BBC NEWS (Mar. 21, 2018), https://www.bbc.com/news/blogs-trending-43470837 [https://perma.cc/UFL4-RM8N].

[234] Alex Hern, *Facebook Usage Falling After Privacy Scandals, Data Suggests*, THE GUARDIAN (Jun. 20, 2019), https://www.theguardian.com/technology/2019/jun/20/facebook-usage-collapsed-since-scandal-data-shows [https://perma.cc/8MRB-ALN5].

settings tools more accessible, days after the Cambridge Analytica story initially broke.[235]

[85] Third, and more generally, transparency disclosures aid consumers in becoming aware of and taking more active steps in combatting misinformation. Much of the American public believes that misinformation and "fake news" are major threats.[236] Annual reports directed at consumers would inform consumers about these threats. Given the intense interest over the role of deepfakes on online platforms, these transparency reports would continue to help journalists, social media watchdog groups, and consumers take critical looks at the media that is being created and moderated.[237] The

[235] *See* Erin Egan & Ashlie Beringer, *It's Time to Make Our Privacy Tools Easier to Find*, FACEBOOK (Mar. 28, 2019), https://about.fb.com/news/2018/03/privacy-shortcuts/ [https://perma.cc/4LD6-CANA] (describing new tools to access data settings, privacy settings, and the ability to download and delete your Facebook data); Carol Cadwalladr & Emma Graham-Harrison, *Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach*, THE GUARDIAN (Mar. 17, 2018), https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election [https://perma.cc/CBC6-3TRV]; *see also* Sam Meredith, *Facebook-Cambridge Analytica: A Timeline of the Data Hijacking Scandal*, CNBC (Apr. 10, 2018), https://www.cnbc.com/2018/04/10/facebook-cambridge-analytica-a-timeline-of-the-data-hijacking-scandal.html [https://perma.cc/AJ8S-28KP] (providing a helpful timeline of the Cambridge Analytica scandal).

[236] *See* Sabrina Siddiqui, *Half of Americans See Fake News as Bigger Threat Than Terrorism, Study Finds*, THE GUARDIAN (Jun. 7, 2019), https://www.theguardian.com/us-news/2019/jun/06/fake-news-how-misinformation-became-the-new-front-in-us-political-warfare [https://perma.cc/2EPL-8RL8] (stating that 70% off Americans believe "fake news and misinformation" have affected "their confidence in government institutions" and "half of Americans view fake news as a bigger threat to the country than terrorism, illegal immigration, violent crime, or racism" according to a Pew Research Center study).

[237] Disclosure investigations have already forced companies to take steps to moderate their content, in many ways more effectively than Section 230 advocates have asked for. Most recently, Pornhub came under heavy scrutiny after a seething New York Times expose of abusive and illegal content on their website. *See* Nicholas Kristof, *The Children of Pornhub*, NEW YORK TIMES (Dec. 4, 2020), https://www.nytimes.com/2020/12/04/opinion/sunday/pornhub-rape-trafficking.html [https://perma.cc/HT3C-NW6R]. Within four days, Pornhub removed many videos, agreed to new policies to

overview of these disclosures would give a better estimate of the extent of the deepfake threat and how consumers can protect themselves.

[86] One major question that must be answered is which online platforms will be subject to the disclosure requirement. Given the costs of creating these reports and conducting research, there is a risk of harming smaller platforms and businesses. This proposal aims to place higher burdens on companies that have higher deepfake traffic. However, precisely defining which platforms will need to comply will require further research from platform advocates and legislatures. This Article tentatively states one way to limit over-inclusivity is to modify requirements from California's Consumer Privacy Act (CCPA).[238] Companies that fall within one of two categories will be liable:

> *1) Independent Company*
>
> a) A for-profit business;
>
> b) that collects United States consumers' personal information (or such information is collected on their behalf) and determine that purposes and means of processing United States consumers' personal information;
>
> c) that does business in the United States **and**
>
> d) Has at *least $25 million in annual gross revenues OR buys, sells, shares, and/or receives the personal information of at*

keep nonconsensual videos off the site and restrict uploads to verified content partners. *See* Russell Brandom, *Pornhub limits uploads and disables downloads after New York Times expose*, THE VERGE (Dec. 8, 2020) https://www.theverge.com/2020/12/8/22164031/pornhub-upload-limit-blocked-download-nyt-kristof-child-abuse [https://perma.cc/AJA2-QK3A]. Just two days after Pornhub's announcement, Visa and Mastercard announced they would cut ties with the site and block customers from using their credit cards to make purchases on the site. *See* Associated Press, *Pornhub: Mastercard and Visa to block use of cards on site after child abuse allegations*, THE GUARDIAN (Dec. 10, 2020), https://www.theguardian.com/us-news/2020/dec/10/pornhub-mastercard-visa-rape-child-abuse-images [https://perma.cc/TDH3-A8RP].

[238] *See* California Consumer Privacy Act of 2018, S. 1121 at §1798.130, 2018 Cal. Sen. (2018).

> *least 200,000 United States consumers, households or devices per year.*
> **OR**
> *2) Controlled Company*
> a) You are a company that **controls or is controlled** by an entity that meets the above criteria and shares **common branding** with that entity.

[87] There are several reasons why these modified requirements could work. An independent study created on behalf of the California state legislature found, generally, that firms with more than 250 employees will meet the $25 million threshold and all businesses with 500+ employees would be subject to these laws.[239] Additionally, 37.5% of businesses in the 100-499 employee category would need to comply with the law under the $25 million threshold.[240] Given most small businesses have an average annual revenue under $25 million, they would likely be excluded under this requirement.[241]

[88] However, companies may be liable under the 200,000 United States consumers, households, or devices per year requirement. The independent study found that any firm that collects personal information from more than 137 consumers or devices a day would meet the CCPA's 50,000 consumers threshold.[242] To alleviate the burden on smaller companies, and given the

[239] David Roland-Holst et. al., *Standardized Regulatory Impact Assessment: California Consumer Privacy of 2018 Regulations, Prepared for Attorney General's Office, California Department of Justice* (Aug. 2019), at 20, http://www.dof.ca.gov/Forecasting/Economics/Major_Regulations/Major_Regulations_Table/documents/CCPA_Regulations-SRIA-DOF.pdf [https://perma.cc/3KDH-DGFN].

[240] *Id.*

[241] *See* Anna Attkisson, *How California's Consumer Privacy Act Will Affect Your Business*, BUSINESS NEWS DAILY (Dec. 31, 2019), https://www.businessnewsdaily.com/10960-ccpa-small-business-impact.html [https://perma.cc/8QCV-2ATB].

[242] *See* Roland-Host et al., *supra* note 239, at 20.

national scope of this requirement, this proposal increased this number of consumers.

[89] Additionally, this proposal excludes the CCPA requirement from (d) to include any company that derives "at least 50 percent" of its annual revenue from selling consumers' personal information.[243] The study found that under the CCPA, the 50,000 consumers and 50% annual revenue requirement would reach 50-75% of California businesses that make under $25 million in revenue.[244] By removing one of these requirements and increasing the consumer requirement for the other, this proposal seeks to limit the burden on smaller companies and businesses as much as possible. However, appropriately tailoring the scope of which companies fall under the transparency disclosures requires further analysis and thought.

[90] Additionally, this Article suggests that financial liability for this proposal utilize a tailored approach to create fines. Fines should focus on the size and nature of companies (i.e. Google vs. DuckDuckGo),[245] the nature of the violation (i.e. negligently missing a disclosure deadline vs. intentionally withholding information), and potentially limiting the damages cap for plaintiffs suing for different damages (i.e. a set amount for negligent omission vs. a higher, punitive amount for intentional violations).

[243] California Consumer Privacy Act of 2018 § 1798.140(c)(1)(C).

[244] *See* Roland-Host ET AL., *supra* note 239, at 20–21.

[245] *Compare*, DUCKDUCKGO, https://duckduckgo.com/about [https://perma.cc/8X6T-AEHY] (stating that the current number of employees is 124 with 2.5 billion monthly searches), *with Google Search Statistics*, INTERNET LIVE STATS, https://www.internetlivestats.com/google-search-statistics/ [https://perma.cc/JR8Q-XU95] (showing that Google receives 3.5 billion searches on a daily basis), *and* Alphabet, *Alphabet Announces Fourth Quarter and Fiscal Year 2018 Results* (Feb. 4, 2019) https://abc.xyz/investor/static/pdf/2018Q4_alphabet_earnings_release.pdf?cache=adc3b38 [https://perma.cc/RU68-ENDC] (noting that Alphabet, Google's parent company, employs almost 100,000 people).

By doing so, this may avoid punitive fines for smaller companies that could lead to consolidation or deterrence.[246]

B. Part II: Collaboration with the Government for Research

[91] The second part of this proposal launches an initiative for platforms to optionally collaborate with the federal government to research and combat the spread of malicious deepfakes. The fruits of this research should be made available for smaller companies and businesses to responsibly implement these tools. Given the complexity and variety of these online platforms, this proposal does not seek to create a mandated form of collaboration or liability for failure to collaborate. Instead, the most fruitful attempt at collaboration should be tailored given each company's respective resources, audience, and reach. As companies change in size and audience interaction, these collaborations can also vary over time.

1. Research

[92] Platforms could provide the government valuable research and data. The federal government recently began researching initiatives on deepfakes.[247] Online platforms could aid the government tremendously with their own research. Specific research on deepfakes and ample datasets of deepfake content could provide valuable resources needed to test these

[246] *See* Alec Stapp, *GDPR After One Year: Costs and Unintended Consequences*, TRUTH ON THE MARKET (May 24, 2019), https://truthonthemarket.com/2019/05/24/gdpr-after-one-year-costs-and-unintended-consequences [https://perma.cc/9UWD-6RYZ] (describing the issues with GDPR forcing consolidation).

[247] *See* Deepfake Report Act of 2019, S. 2065, 116th Cong. §3(a) (2019), https://www.congress.gov/116/bills/s2065/BILLS-116s2065es.pdf [https://perma.cc/6H8X-E4YH] (requiring assessments on technologies used to create forgeries, descriptions of these forgeries, and the use of these forgeries by non-governmental entities).

technologies in real time.[248] Sharing information also provides datasets that are constantly evolving and providing the most current deepfakes that are being used. The research collaboration between the government and private companies could provide valuable metrics to combat deepfakes.[249] This collaboration could create better ways of assessing how deepfakes have evolved in order to create better policy responses "according to the technology's contemporary performance as well as its likely evolution."[250] These kinds of metrics could also help create more cohesion among private companies and various sectors in their conversation about deepfakes. Creating a shared idea of what deepfakes entail, or how they are defined, could help unify a fragmented field.[251] These metrics could provide valuable sector-wide analysis of deepfakes and which solutions provide the best results in certain contexts.[252]

[93] For a generation that grew up with the Patriot Act and Cambridge-Analytica scandal in its rearview mirror, the thought of online platforms and

[248] *See* Nick Dufour & Andrew Gully, *Contributing Data to Deepfake Detection Research*, GOOGLE: AI BLOG (Sept. 24, 2019), https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html [https://perma.cc/5MXC-W5AB] (releasing datasets to support the "thriving research community around mitigating potential harms from misuses of synthetic data"); *Creating a Dataset and a Challenge for Deepfakes*, FACEBOOK: FACEBOOK AI (Sept. 5, 2019), https://ai.facebook.com/blog/deepfake-detection-challenge/ [https://perma.cc/8NBK-Z3EQ] (describing the need for Facebook to create a dataset given the importance for data that "is freely available for the community to use" and "realistic").

[249] *Hearing on The National Security Challenges of Artificial Intelligence*, *supra* note 33, at 9.

[250] *Id.*

[251] *See supra* text accompanying notes 60–67.

[252] *See* SELECT COMM. ON ARTIFICIAL INTELLIGENCE, NAT'L SCI. & TECH. COUNCIL, THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN: 2019 UPDATE 30, 35 (2019), https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf [https://perma.cc/NRS2-44WC] (discussing the need for developing "open-source software libraries and toolkits").

the federal government sharing data can be terrifying. However, actions can be taken to reduce consumer concerns over data collection and sharing. Online platforms that share data can anonymize user-profiles and tracking methodologies. If a malicious deepfake was shared from a specific user's profile, the platform could provide other relevant characteristics about the user's audience, their general platform usage, and the content itself rather than the user's exact identity. Also, shared data, and the extent of this shared data, could be information provided to consumers. This provides transparency around the type of information platforms are sharing and potentially give consumers the ability to shape this information.

[94] Additionally, many platforms have shown a general reluctance to share data until needed, which could provide a necessary "check" on giving extraneous information. For example, Facebook generally requires court orders, subpoenas, or warrants to grant specific user information.[253] Additionally, Google states that under the Fourth Amendment and ECPA, the United States government is very limited in its ability to collect user data outside court orders, warrants, and subpoenas.[254] As a result, online platforms will have the freedom and the ability to narrowly tailor the provided data and work with their users to protect their privacy.

2. Technology Development

[95] Independently, most online platforms and the government have begun their own initiatives to better detect deepfakes. Online platforms are now working collaboratively to create competitions that develop better deepfake detection methods. For example, Amazon, Microsoft, Facebook, and the Partnership on AI created a new Deepfake Detection Challenge for contestants to "build better detection tools" for the "technically demanding

[253] *See Information for Law Enforcement Authorities*, FACEBOOK, https://www.facebook.com/safety/groups/law/guidelines/ [https://perma.cc/F5D7-TSM8].

[254] *See How Google Handles Government Requests for user Information*, GOOGLE, https://policies.google.com/terms/information-requests [https://perma.cc/DN6W-UKYA].

and rapidly evolving challenge" of deepfakes.[255] Prizes range from $40,000 to $500,000, with users using a closed dataset provided by the platforms.[256]

[96] The government has two programs under DARPA, or the Defense Advanced Research Projects Agency, that specifically handles deepfakes.[257] Media Forensics or MediFor, is being used to develop algorithms that "automatically assess the integrity of photos and videos and to provide analysts with information about how counterfeit content was generated."[258] Semantics Forensics, or SemaFor, works separately to "develop algorithms that will automatically detect, attribute, and characterize . . . various types of deep fakes" as benign or malicious.[259] In 2018, the federal government issued their own deepfake detection contest under MediFor.[260] In 2020, the federal government authorized five million dollars to the IARPA, or the Intelligence Advanced Research Projects Activity, to host a competition to develop new deepfake detection tools.[261]

---

[255] Maithreyan Surya, *The Decade of Artificial Intelligence*, TOWARDS DATA SCIENCE, https://towardsdatascience.com/the-decade-of-artificial-intelligence-6fcaf2fae473 [https://perma.cc/9PFV-9VU5].

[256] *Deepfake Detection Challenge*, KAGGLE, https://www.kaggle.com/c/deepfake-detection-challenge/overview/prizes [https://perma.cc/7EBM-7HM5].

[257] KELLEY M. SAYLER & LAURIE A. HARRIS, DEEP FAKES AND NATIONAL SECURITY 1, CONGRESSIONAL RESEARCH SERVICE (Oct. 14, 2019), https://crsreports.congress.gov/product/pdf/IF/IF11333 [https://perma.cc/5BAS-6X3A].

[258] *Id.* at 1–2.

[259] *Id.* at 2.

[260] *See Media Forensics Challenge 2018*, NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY (Dec. 10, 2019), https://www.nist.gov/itl/iad/mig/media-forensics-challenge-2018 [https://perma.cc/9YHQ-62CT].

[261] Alex Engler, *Fighting Deepfakes When Detection Fails*, BROOKINGS INSTITUTE (Nov. 14, 2019), https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/#footnote-17 [https://perma.cc/5SGY-CVP4].

[97] Combining these efforts together could create powerful detection tools for both the private sector and the government.[262] By working with the government, private companies could more effectively share and analyze new technology. For example, if one company was able to successfully detect a specific type of deepfake video, they could share these results and methodologies with other stakeholders. Additionally, this collaboration could provide more funding for initiatives like deepfake detection competitions to reach a broader audience.

3. Dissemination of Research Safely and Responsibly

[98] The government and private companies should release the fruits of their research, specifically new detection technologies, in a controlled and responsible manner.[263] This proposal adopts the model used by OpenAI, an artificial intelligence research company in San Francisco, CA, whose primary purpose is to ensure AI "benefits all of humanity."[264]

[99] OpenAI's dissemination model uses two release strategies to ensure AI technology is being distributed responsibly.[265] First, OpenAI utilizes

[262] *See The National Artificial Intelligence Research and Development Strategic Plan: 2019 Update*, NATIONAL SCIENCE & TECHNOLOGY COUNCIL at 42 (June 2019), https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf [https://perma.cc/66YJ-XM9S] (describing various ways the federal government works with the private sector to develop new technologies and resources, such as the Silicon Valley Doctorate Program or the "Top Health" Tech Spring Initiative).

[263] *See Artificial Intelligence for the American People*, THE WHITE HOUSE, https://trumpwhitehouse.archives.gov/ai/executive-order-ai/ [https://perma.cc/7BKN-VCN8] (describing the multiple actions taken by the Executive Office, including an "AI Executive Order" and various research initiatives to promote the development of new AI technology and research).

[264] *Hearing on "The National Security Challenges of Artificial Intelligence, Manipulated Media, and 'Deep Fakes'*" *supra* note 33, at 2.

[265] *See id.* at 5–6.

"Staged Releases." [266] With Staged Releases, the company releases a "small version" of their AI technology, then releases a "medium" version of the technology three months later.[267] The difference between these technologies is a difference in parameters and performance at tasks.[268] Staged Releases allow OpenAI to "slowly introduce a technology into the world . . . better monitor its usage and diffusion," and to allow them to better measure and "calibrate" their "own threat model and systems of analysis" when it comes to releasing AI technology.[269]

[100] Second, OpenAI discussed using "Partnership" release strategies.[270] Under a Partnership release strategy, Open AI would "privately and non-commercially partner with other companies, institutions, and academia research groups" to share their technology in order to conduct better "research into mitigations and threat models and technical interventions."[271] OpenAI believes that, by utilizing these release strategies, it can be "more thoughtful" in its release of AI technology and can prevent "potentially abusive" uses.[272]

[101] This proposal advocates for the government and private companies to share their technological and research developments with other stakeholders using these release strategies. Detection technologies must be available for smaller companies, journalist organizations, developers, and

[266] *Id.*

[267] *Id.*

[268] *See id.* at 6.

[269] *Hearing on "The National Security Challenges of Artificial Intelligence, Manipulated Media, and 'Deep Fakes'" supra* note 33, at 6.

[270] *Id.*

[271] *Id.*

[272] *Id.*

even the public to truly prevent the spread of deepfakes.[273] New technologies should first be shared in a limited capacity, via a modified partnership release, with other companies participating in government collaborations. This initial partnership release would test the technology, create performance metrics, and provide feedback.

[102] Once the technology has completed its initial partnership release, it should be distributed in a staged release to outside stakeholders. In this staged release, companies and the government could license detection technology to users, entities, and organizations that have been properly verified and vetted to ensure the technology is not used maliciously. Additionally, software monitoring could ensure that the technology is not distributed freely to improper parties. At its most extreme, the technology may have a staged release to the public. This would likely be done in exchange for the ability to gather data about the software from users in order to create larger deepfake datasets. However, much like how OpenAI would release "smaller" versions of its software, these public releases would utilize limited versions of the software. By utilizing these systems, the government and private sector collaboration empowers and gives tools to all stakeholders.[274]

[273] *See* John Bowers et al., *What Should Newsrooms Do About Deepfakes? These Three Things, For Starters*, NIEMAN LAB (Nov. 20, 2019), https://www.niemanlab.org/2019/11/what-should-newsrooms-do-about-deepfakes-these-three-things-for-starters [https://perma.cc/Y4MR-Y9GQ] (describing how journalists are implicated and must take proactive steps to combat deepfakes).

[274] Miles Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, FUTURE OF HUMANITY INSTITUTE (Feb. 2018), https://img1.wsimg.com/blobby/go/3d82daa4-97fe-4096-9c6b-376b92c619de/downloads/MaliciousUseofAI.pdf?ver=1553030594217 [https://perma.cc/E4JB-A798] (stressing the need to have stakeholder input in the development, distribution, and implementation of AI technologies).

C. Part III: Investing in Public Education

[103] The government and platforms should invest in free, accessible resources that train the public about deceptive manipulated media. These resources should be distributed and updated as technology evolves.[275] Social media platforms could provide brief, mandatory detection lessons for their users. One could imagine Facebook asking its users to complete a short quiz regarding manipulated media, providing an assessment of the user's answers, and reviewing the platform's respective policies on manipulated media and content takedown.

[104] The U.S. government should create programs and fund efforts to promote media literacy. A 2019 study found there was a "lack of understanding about what it means to be "'media literate,'" with a need for more "broad-based funding" to support "media literacy education and encourage high quality scalable media literacy programs" in the United States.[276] The United States' "media education goals are far from adequately being met,"[277] and have seemingly lagged behind its peers historically.[278]

---

[275] *See Digital Literacy Library*, FACEBOOK, https://www.facebook.com/safety/educators [https://perma.cc/E9UU-3U47] (describing Facebook providing some educational resources that give lessons about media literacy, primarily targeted to children and teens).

[276] Sherri Culver & Theresa Redmond, *Snapshot 2019: The State of Media Literacy Education in the U.S.*, NATIONAL ASSOCIATION FOR MEDIA LITERACY EDUCATION (2019), https://namle.net/wp-content/uploads/2020/10/SOML_FINAL.pdf [https://perma.cc/9P5U-MUJC].

[277] *Id.*

[278] *See Media Literacy in the USA*, CENTER FOR MEDIA LITERACY, https://www.medialit.org/reading-room/media-literacy-usa [https://perma.cc/V69C-8EXB] (during the 1980s and 1990s, the field of media literacy was "flourishing"). *But see* Laura Lederer, *What Are Other Countries Doing in Media Education?*, CENTER FOR MEDIA LITERACY (1988), https://www.medialit.org/reading-room/what-are-other-countries-doing-media-education [https://perma.cc/8U2Q-XC6A] (noting that the United States was lagging "behind these countries.").

[105] The U.S. government must provide resources to train the public on digital literacy and media manipulation.[279] The Finish government sponsored a free course that was developed alongside the University of Helsinki and several companies to familiarize individuals with developing AI technologies.[280] The U.S. government could create similar courses for individuals to better educate themselves about these issues. Additionally, the government should "establish an online, central repository for the collection, curation, and aggregation of resources" and lessons for a "variety of ages, grades, and contexts" to improve media literacy.[281] These resources would be central in educating other stakeholders, such as journalists and fact-checkers, on how to effectively use new detection technologies in their own fields.[282]

[106] A better educated consumer would mitigate the problem of the "liar's dividend."[283] The "liar's dividend" theory posits that malicious deepfakes could create consumers who are distrustful of both fake media *and* truthful media.[284] Given a constant state of misinformation, "liars aiming to dodge responsibility for their real words and actions will become

[279] *See* Digital Citizenship and Media Literacy Act, S. 2240, 116th Cong. § 3(b) (2019).

[280] *Hearing on The National Security Challenges of Artificial Intelligence*, *supra* note 33, at 10 (arguing that the government should invest in comprehensive AI education).

[281] *See, e.g.*, Culver & Redmond, *supra* note 276, at 10; *see also* Joint Communication to the European Parliament, The European Council, The Council, The European Economic and Social Committee and the Committee of the Regions - Action Plan Against Disinformation (EC) at 10 (May 12, 2018), https://eeas.europa.eu/sites/eeas/files/action_plan_against_disinformation.pdf [https://perma.cc/JK6Y-7KQJ] (identifying raising "public awareness" and "media literacy to empower Union citizens to better identify and deal with disinformation" as part of its strategy to fight disinformation).

[282] Engler, *supra* note 261.

[283] Chesney & Citron, *supra* note 2, at 1785.

[284] *Id.*

more credible."[285] In short, so many individuals will start to cry wolf that everything seems false.

[107] Investing in public education about deepfakes and media literacy can lessen this effect.[286] Rather than being skeptical of all content, educational resources will train the public to be more critical upfront. This critical nature of analyzing content, discussing its veracity, and attempting to discover if it was maliciously manipulated is core to creating more democratic discourse.[287] Rather than having a population that gullibly

[285] *Id.*

[286] *See, e.g.*, Ullrich K.H. Ecker et. al., *Explicit Warnings Reduce but Do Not Eliminate the Continued Influence of Misinformation*, 38(8) Memory & Cognition 1087, 1094, 1096 (2010) (finding that warnings on misinformation did help reduce reliance on misinformation, albeit not to a substantive amount, but educating people about the negative continued reliance on misinformation could help lessen this effect); Andrew M. Guess et al., *A Digital Media Literacy Intervention Increases Discernment Between Mainstream and False News in the United States and India*, 117 Proceedings of the National Academy of Sciences 15536, 15537, 15541 (2020), https://www.pnas.org/content/pnas/117/27/15536.full.pdf [https://perma.cc/ZZQ3-TDYB] (finding that providing "interventions" of consumers using media literacy guidelines, reduced the negative effects of false headlines, increased discernment of mainstream news, with "no measurable decrease in the perceived accuracy of mainstream news headlines,"); Monica Bulger & Patrick Davison, *The Promises, Challenges, and Futures of Media Literacy*, 10 J. OF MEDIA LITERACY EDUC. 1, 8, 10 (2018), https://digitalcommons.uri.edu/cgi/viewcontent.cgi?article=1365&context=jmle [https://perma.cc/ZSM3-64TF] (stating that media literacy "could increase critical approaches to media, an appreciation that people approach media differently, and a recognition of the effects of violence in media," but acknowledging that media literacy is not a silver bullet and has its shortcomings, specifically gaps in information about the commercial sources of information and using outdated metrics of assessing accuracy).

[287] *See* Saoirse De Paor & Bahareh Heravi, *Information Literacy and Fake News: How the Field of Librarianship Can Help Combat the Epidemic of Fake News*, 46 J. OF ACAD. LIBRARIANSHIP at 5 (2020) (explaining how librarians promote and develop educational programs teaching students how to assess fake news media).

believes or disbelieves everything, media literacy can create a more informed online body that approaches media and content with caution.[288]

D. This Proposal Best Addresses Deepfake Harms

[108] This proposal best addresses deepfake harms while protecting the principles of innovation, expression, and statutory protections for online speech. Additionally, this proposal counters the three unique harms posed by deepfakes. First, it negates the effects of upfront harm. Increased collaboration among the government and online platforms could develop powerful new technologies that could provide better detection of malicious deepfakes earlier. However, and perhaps more importantly, transparency disclosures and an investment in public media literacy would make consumers more aware of manipulated content. This could lead more consumers to be critical of potentially manipulated content and reduce the upfront belief of a deepfake.

[109] Second, this proposal addresses the issue of content spreading rapidly throughout social media. Once again, all three parts of this proposal aid in mitigating this harm. A requirement of disclosures and public education will make consumers aware that malicious deepfake content can and may spread. This factor demonstrates the power of "flipping" the "liar's

[288] *See, e.g.*, John Cook et. al., *Misinformation and How to Correct It*, EMERGING TRENDS IN THE SOCIAL AND BEHAVIORAL SCIENCES at 6 (2015), https://www.emc-lab.org/uploads/1/1/3/6/113627673/cookecker.2015.etsbs.pdf [https://perma.cc/2A4L-HJPL] (noting that teaching the ability to refute misinformation in classrooms "can be an opportunity to foster critical thinking," and encourage students to "skeptically assess empirical evidence and draw valid conclusions from the evidence"); S. Mo Jones-Jang et.al., *Does Media Literacy Help Identification of Fake News? Information Literacy Helps, but Other Literacies Don't*, AM. BEHAV. SCIENTIST at 12 (2019), https://www.researchgate.net/publication/335352499_Does_Media_Literacy_Help_Identification_of_Fake_News_Information_Literacy_Helps_but_Other_Literacies_Don't [https://perma.cc/C69H-K8Y4] (referencing studies that found that teachings should incorporate a variety of different literacy styles, but emphasis should be placed on giving users "the skills and competencies to sustain and update their access to rapidly changing information systems," like locating "fact-checking websites or relevant online tools efficiently" and evaluating "multiple sources").

dividend." As more consumers become critical of the media they consume, they hasten the spread of malicious viral content. Or, if the content spreads, more consumers may be quick to label and call out the content as fake, satire, or a parody. Investment in collaboration between the government and online platforms may also mitigate this spread. As better datasets are developed and online platforms work in unison, better technologies can be developed. This may include faster recognition, labelling potentially malicious content, and creating faster algorithm responses to prevent the platform from emphasizing the content on users' feeds.

[110] Third, this proposal best addresses the issue of deepfakes escaping strict definition and constantly evolving. Given the focus on education and innovation, this proposal fosters an experimentation and open communication about the efficiency of these solutions among the key stakeholders. By fostering these principles, it provides more flexible solutions that allow for a deeper analysis of the deepfake field.

[111] This proposal will not lead to perfect technologies that will automatically detect deepfakes today. However, investing in the right tools today can lead to better, more permanent solutions. Rather than creating overly broad legislation, this proposal urges platforms and the government to collaborate and take intentional, researched action to address deepfakes.

VI. CONCLUSION

[112] Deepfakes are a new frontier. Given the rate at which deepfake technology is becoming more accessible and more advanced, deepfakes will continue to be a tool that is easily accessible and utilized by many. However, ensuring the spread of deepfakes aligns with safe uses of this new technology is crucial. As this Article explored, deepfakes can be used maliciously to destructive ends against both individuals and society. To mitigate these harms, this Article puts forth a new proposal for increased transparency from online platforms, collaboration between platforms and the government, and investment in educational resources to improve media literacy more generally.