Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security

Bobby Chesney* and Danielle Citron**

Harmful lies are nothing new. But the ability to distort reality has taken an exponential leap forward with "deep fake" technology. This capability makes it possible to create audio and video of real people saying and doing things they never said or did. Machine learning techniques are escalating the technology's sophistication, making deep fakes ever more realistic and increasingly resistant to detection. Deep-fake technology has characteristics that enable rapid and widespread diffusion, putting it into the hands of both sophisticated and unsophisticated actors.

While deep-fake technology will bring certain benefits, it also will introduce many harms. The marketplace of ideas already suffers from truth decay as our networked information environment interacts in toxic ways with our cognitive biases. Deep fakes will exacerbate this problem significantly. Individuals and businesses will face novel forms of exploitation, intimidation, and personal sabotage. The risks to our democracy and to national security are profound as well.

Our aim is to provide the first in-depth assessment of the causes and consequences of this disruptive technological change, and to explore the existing and potential tools for responding to it. We survey a broad array of responses, including: the role of technological solutions; criminal penalties, civil liability, and regulatory action; military and covert-action responses; economic sanctions; and market developments. We cover the waterfront from immunities to immutable authentication trails, offering recommendations to improve law and policy and anticipating the pitfalls embedded in various solutions.

INTRODUCTION

Through the magic of social media, it all went viral: a vivid photograph, an inflammatory fake version, an animation expanding on the fake, posts debunking the fakes, and stories trying to make sense of the situation.[1] It was both a sign of the times and a cautionary tale about the challenges ahead.

The episode centered on Emma González, a student who survived the horrific shooting at Marjory Stoneman Douglas High School in Parkland, Florida, in February 2018. In the aftermath of the shooting, a number of the students emerged as potent voices in the national debate over gun control. Emma, in particular, gained prominence thanks to the closing speech she delivered during the "March for Our Lives" protest in Washington, D.C., as well as a contemporaneous article she wrote for *Teen Vogue*.[2] Fatefully, the *Teen Vogue*

---

1.   Alex Horton, *A Fake Photo of Emma González Went Viral on the Far Right, Where Parkland Teens are Villains*, WASH. POST (Mar. 26, 2018), https://www.washingtonpost.com/news/the-intersect/wp/2018/03/25/a-fake-photo-of-emma-gonzalez-went-viral-on-the-far-right-where-parkland-teens-are-villains/?utm_term=.0b0f8655530d [https://perma.cc/6NDJ-WADV].

2.   *Florida Student Emma Gonzalez* [sic] *to Lawmakers and Gun Advocates: 'We call BS'*, CNN (Feb. 17, 2018), https://www.cnn.com/2018/02/17/us/florida-student-emma-gonzalez-speech/index.html [https://perma.cc/ZE3B-MVPD]; Emma González, *Emma González on Why This Generation Needs Gun Control*, TEEN VOGUE (Mar. 23, 2018), https://www.teenvogue.com/story/emma-gonzalez-parkland-gun-control-cover?mbid=social_twitter [https://perma.cc/P8TQ-P2ZR].

piece incorporated a video entitled "This Is Why We March," including a visually arresting sequence in which Emma rips up a large sheet displaying a bullseye target.

A powerful still image of Emma ripping up the bullseye target began to circulate on the Internet. But soon someone generated a fake version, in which the torn sheet is not a bullseye, but rather a copy of the Constitution of the United States. While on some level the fake image might be construed as artistic fiction highlighting the inconsistency of gun control with the Second Amendment, the fake was not framed that way. Instead, it was depicted as a true image of Emma González ripping up the Constitution.

The image soon went viral. A fake of the video also appeared, though it

was more obvious that it had been manipulated. Still, the video circulated widely, thanks in part to actor Adam Baldwin circulating it to a quarter million followers on Twitter (along with the disturbing hashtag #Vorwärts—the German word for "forward," a reference to neo-Nazis' nod to the word's role in a Hitler Youth anthem). [3]

Several factors combined to limit the harm from this fakery. First, the genuine image already was in wide circulation and available at its original source. This made it fast and easy to fact-check the fakes. Second, the intense national attention associated with the post-Parkland gun control debate and, especially, the role of students like Emma in that debate, ensured that journalists paid attention to the issue, spending time and effort to debunk the fakes. Third, the fakes were of poor quality (though audiences inclined to believe their message might disregard the red flags).

Even with those constraints, though, many believed the fakes, and harm ensued. Our national dialogue on gun control has suffered some degree of

---

3.   *See* Horton, *supra* note 1.

distortion; Emma has likely suffered some degree of anguish over the episode; and other Parkland victims likely felt maligned and discredited. Falsified imagery, in short, has already exacted significant costs for individuals and society. But the situation is about to get much worse, as this Article shows.

Technologies for altering images, video, or audio (or even creating them from scratch) in ways that are highly -realistic and difficult to detect are maturing rapidly. As they ripen and diffuse, the problems illustrated by the Emma González episode will expand and generate significant policy and legal challenges. Imagine a deep fake video, released the day before an election, making it appear that a candidate for office has made an inflammatory statement. Or what if, in the wake of the Trump-Putin tête-à-tête at Helsinki in 2018, someone circulated a deep fake audio recording that seemed to portray President Trump as promising not to take any action should Russia interfere with certain NATO allies. Screenwriters are already building such prospects into their plotlines.[4] The real world will not lag far behind.

Pornographers have been early adopters of the technology, interposing the faces of celebrities into sex videos. This has given rise to the label "deep fake" for such digitized impersonations. We use that label here more broadly, as shorthand for the full range of hyper-realistic digital falsification of images, video, and audio.

This full range will entail, sooner rather than later, a disturbing array of malicious uses. We are by no means the first to observe that deep fakes will migrate far beyond the pornography context, with great potential for harm.[5] We

4.   *See, e.g.*, Vindu Goel & Sheera Frenkel, *In India Election, False Posts and Hate Speech Flummox Facebook*, N. Y. TIMES (Apr. 1, 2019), https://www.nytimes.com/2019/04/01/technology/india-elections-facebook.html [https://perma.cc/B9CP-MPPK**]** (describing the deluge of fake and manipulated videos and images circulated in the lead up to elections in India); *Homeland: Like Bad at Things* (Showtime television broadcast Mar. 4, 2018), https://www.sho.com/homeland/season/7/episode/4/like-bad-at-things [https://perma.cc/25XK-NN3Y]; *Taken: Verum Nocet* (NBC television broadcast Mar. 30, 2018) https://www.nbc.com/taken/video/verum-nocet/3688929 [https://perma.cc/CVP2-PNXZ] (depicting a deep-fake video in which a character appears to recite song lyrics); *The Good Fight: Day 408* (CBS television broadcast Mar. 4, 2018) (depicting fake audio purporting to be President Trump); *The Good Fight: Day 464* (CBS television broadcast Apr. 29, 2018) (featuring a deep-fake video of the alleged "golden shower" incident involving President Trump).

5.   *See, e.g.*, Samantha Cole, *We Are Truly Fucked: Everyone is Making AI-Generated Fake Porn Now*, VICE: MOTHERBOARD (Jan. 24, 2018), https://motherboard.vice.com/en_us/article/bjye8a/reddit-fake-porn-app-daisy-ridley [https://perma.cc/V9NT-CBW8**]** ("[T]echnology[] allows anyone with sufficient raw footage to work with to convincingly place any face in any video."); *see also* @BuzzFeed, *You Won't Believe What Obama Says in This Video*, TWITTER (Apr. 17, 2018, 8:00 AM), https://twitter.com/BuzzFeed/status/986257991799222272 [https://perma.cc/C38K-B377] ("We're entering an era in which our enemies can make anyone say anything at any point in time."); Tim Mak, *All Things Considered: Technologies to Create Fake Audio and Video Are Quickly Evolving*, NAT'L PUB. RADIO (Apr. 2, 2018), https://www.npr.org/2018/04/02/598916380/technologies-to-create-fake-audio-and-video-are-quickly-evolving [https://perma.cc/NY23-YVQD] (discussing deep-fake videos created for political reasons and misinformation campaigns); Julian Sanchez (@normative), TWITTER (Jan. 24, 2018, 12:26 PM) ("The prospect of any Internet rando being able to swap anyone's face into

do, however, provide the first comprehensive survey of these harms and potential responses to them. We break new ground by giving early warning regarding the powerful incentives that deep fakes produce for privacy-destructive solutions.

This Article unfolds as follows. Part I begins with a description of the technological innovations pushing deep fakes into the realm of hyper-realism and making them increasingly difficult to debunk. It then discusses the amplifying power of social media and the confounding influence of cognitive biases.

Part II surveys the benefits and the costs of deep fakes. The upsides of deep fakes include artistic exploration and educative contributions. The downsides of deep fakes, however, are as varied as they are costly. Some harms are suffered by individuals or groups, such as when deep fakes are deployed to exploit or sabotage individual identities and corporate opportunities. Others impact society more broadly, such as distortion of policy debates, manipulation of elections, erosion of trust in institutions, exacerbation of social divisions, damage to national security, and disruption of international relations. And, in what we call the "liar's dividend," deep fakes make it easier for liars to avoid accountability for things that are in fact true.

Part III turns to the question of remedies. We survey an array of existing or potential solutions involving civil and criminal liability, agency regulation, and "active measures" in special contexts like armed conflict and covert action. We also discuss technology-driven market responses, including not just the promotion of debunking technologies, but also the prospect of an alibi service, such as privacy-destructive life logging. We find, in the end, that there are no silver-bullet solutions. Thus, we couple our recommendations with warnings to the public, policymakers, and educators.

I.
TECHNOLOGICAL FOUNDATIONS OF THE DEEP-FAKES PROBLEM

Digital impersonation is increasingly realistic and convincing. Deep-fake technology is the cutting-edge of that trend. It leverages machine-learning algorithms to insert faces and voices into video and audio recordings of actual people and enables the creation of realistic impersonations out of digital whole cloth.[6] The end result is realistic-looking video or audio making it appear that someone said or did something. Although deep fakes can be created with the consent of people being featured, more often they will be created without it. This Part describes the technology and the forces ensuring its diffusion, virality, and entrenchment.

porn is incredibly creepy. But my first thought is that we have not even scratched the surface of how bad 'fake news' is going to get.").

    6.   *See* Cole, *supra* note 5.

A. *Emergent Technology for Robust Deep Fakes*

Doctored imagery is neither new nor rare. Innocuous doctoring of images—such as tweaks to lighting or the application of a filter to improve image quality—is ubiquitous. Tools like Photoshop enable images to be tweaked in both superficial and substantive ways.[7] The field of digital forensics has been grappling with the challenge of detecting digital alterations for some time.[8] Generally, forensic techniques are automated and thus less dependent on the human eye to spot discrepancies.[9] While the detection of doctored audio and video was once fairly straightforward,[10] the emergence of generative technology capitalizing on machine learning promises to shift this balance. It will enable the production of altered (or even wholly invented) images, videos, and audios that are more realistic and more difficult to debunk than they have been in the past. This technology often involves the use of a "neural network" for machine learning. The neural network begins as a kind of tabula rasa featuring a nodal network controlled by a set of numerical standards set at random.[11] Much as experience refines the brain's neural nodes, examples train the neural network system.[12] If the network processes a broad array of training examples, it should be able to create increasingly accurate models.[13] It is through this process that neural networks categorize audio, video, or images and generate realistic impersonations or alterations.[14]

---

7.   *See, e.g.*, Stan Horaczek, *Spot Faked Photos Using Digital Forensic Techniques*, POPULAR SCIENCE (July 21, 2017), https://www.popsci.com/use-photo-forensics-to-spot-faked-images [https://perma.cc/G72B-VLF2] (depicting and discussing a series of manipulated photographs).

8.   Doctored images have been prevalent since the advent of the photography. *See* PHOTO TAMPERING THROUGHOUT HISTORY, http://pth.izitru.com [https://perma.cc/5QSZ-NULR]. The gallery was curated by FourandSix Technologies, Inc.

9.   *See* Tiffanie Wen, *The Hidden Signs That Can Reveal a Fake Photo*, BBC FUTURE (June 30, 2017), http://www.bbc.com/future/story/20170629-the-hidden-signs-that-can-reveal-if-a-photo-is-fake [https://perma.cc/W9NX-XGKJ]. IZITRU.COM was a project spearheaded by Dartmouth's Dr. Hany Farid. It allowed users to upload photos to determine if they were fakes. The service was aimed at "legions of citizen journalists who want[ed] to dispel doubts that what they [were] posting [wa]s real." Rick Gladstone, *Photos Trusted but Verified*, N.Y. TIMES (May 7, 2014), https://lens.blogs.nytimes.com/2014/05/07/photos-trusted-but-verified [https://perma.cc/7A73-URKP].

10.   *See* Steven Melendez, *How DARPA's Fighting Deepfakes*, FAST COMPANY (Apr. 4, 2018), https://www.fastcompany.com/40551971/can-new-forensic-tech-win-war-on-ai-generated-fake-images [https://perma.cc/9A8L-LFTQ].

11.   Larry Hardesty, *Explained: Neural Networks,* MIT NEWS (Apr. 14, 2017), http://news.mit.edu/2017/explained-neural-networks-deep-learning-0414 [https://perma.cc/VTA6-4Z2D].

12.   Natalie Wolchover, *New Theory Cracks Open the Black Box of Deep Neural Networks,* WIRED (Oct. 8, 2017), https://www.wired.com/story/new-theory-deep-learning [https://perma.cc/UEL5-69ND].

13.   Will Knight, *Meet the Fake Celebrities Dreamed Up By AI*, MIT TECH. REV. (Oct. 31, 2017), https://www.technologyreview.com/the-download/609290/meet-the-fake-celebrities-dreamed-up-by-ai [https://perma.cc/D3A3-JFY4].

14.   Will Knight, *Real or Fake? AI is Making it Very Hard to Know,* MIT TECH. REV. (May 1, 2017), https://www.technologyreview.com/s/604270/real-or-fake-ai-is-making-it-very-hard-to-know [https://perma.cc/3MQN-A4VH].

To take a prominent example, researchers at the University of Washington have created a neural network tool that alters videos so speakers say something different from what they originally said.15 They demonstrated the technology with a video of former President Barack Obama (for whom plentiful video footage was available to train the network) that made it appear that he said things that he had not.16

By itself, the emergence of machine learning through neural network methods would portend a significant increase in the capacity to create false images, videos, and audio. But the story does not end there. Enter "generative adversarial networks," otherwise known as GANs. The GAN approach, invented by Google researcher Ian Goodfellow, brings two neural networks to bear simultaneously.17 One network, known as the generator, draws on a dataset to produce a sample that mimics the dataset.18 The other network, the discriminator, assesses the degree to which the generator succeeded.19 In an iterative fashion, the assessments from the discriminator inform the assessments of the generator. The result far exceeds the speed, scale, and nuance of what human reviewers could achieve.20 Growing sophistication of the GAN approach is sure to lead to the production of increasingly convincing deep fakes.21

---

15. SUPASORN SUWAJANAKORN ET AL., SYNTHESIZING OBAMA: LEARNING LIP SYNC FROM AUDIO, 36 ACM TRANSACTIONS ON GRAPHICS, no. 4, art. 95 (July 2017), http://grail.cs.washington.edu/projects/AudioToObama/siggraph17_obama.pdf [https://perma.cc/7DCY-XK58]; James Vincent, *New AI Research Makes It Easier to Create Fake Footage of Someone Speaking*, VERGE (July 12, 2017), https://www.theverge.com/2017/7/12/15957844/ai-fake-video-audio-speech-obama [https://perma.cc/3SKP-EKGT].

16. Charles Q. Choi*, AI Creates Fake Obama,* IEEE SPECTRUM (July 12, 2017), https://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/ai-creates-fake-obama [https://perma.cc/M6GP-TNZ4]; *see also* Joon Son Chung et al., *You Said That?* (July 18, 2017) (British Machine Vision conference paper), https://arxiv.org/abs/1705.02966 [https://perma.cc/6NAH-MAYL].

17. *See* Ian J. Goodfellow et al., Generative Adversarial Nets (June 10, 2014) (Neural Information Processing Systems conference paper), https://arxiv.org/abs/1406.2661 [https://perma.cc/97SH-H7DD] (introducing the GAN approach); *see also* Tero Karras, et al., Progressive Growing of GANs for Improved Quality, Stability, and Variation, ICLR 2018, at 1-2 (Apr. 2018) (conference paper), http://research.nvidia.com/sites/default/files/pubs/2017-10_Progressive-Growing-of/karras2018iclr-paper.pdf [https://perma.cc/RSK2-NBAE] (explaining neural networks in the GAN approach).

18. Karras, *supra* note 17, at 1.

19. *Id.*

20. *Id.* at 2.

21. Consider research conducted at Nvidia. Karras, *supra* note 17, at 2 (explaining a novel approach that begins training cycles with low-resolution images and gradually shifts to higher-resolution images, producing better and much quicker results). The *New York Times* recently profiled the Nvidia team's work. *See* Cade Metz & Keith Collins, *How an A.I. 'Cat-and-Mouse Game' Generates Believable Fake Photos*, N.Y. TIMES (Jan. 2, 2018), https://www.nytimes.com/interactive/2018/01/02/technology/ai-generated-photos.html [https://perma.cc/6DLQ-RDWD]. For further illustrations of the GAN approach, *see* Martin Arjovsky et al., Wasserstein GAN (Dec. 6, 2017) (unpublished manuscript) (on file with California Law Review); Chris Donahue et al., Semantically Decomposing the Latent Spaces of Generative Adversarial Networks, ICLR 2018 (Feb. 22, 2018) (conference paper) (on file with California Law Review)*,*

The same is true with respect to generating convincing audio fakes. In the past, the primary method of generating audio entailed the creation of a large database of sound fragments from a source, which would then be combined and reordered to generate simulated speech. New approaches promise greater sophistication, including Google DeepMind's "Wavenet" model,[22] Baidu's DeepVoice,[23] and GAN models.[24] Startup Lyrebird has posted short audio clips simulating Barack Obama, Donald Trump, and Hillary Clinton discussing its technology with admiration.[25]

In comparison to private and academic efforts to develop deep-fake technology, less is currently known about governmental research.[26] Given the possible utility of deep-fake techniques for various government purposes— including the need to defend against hostile uses—it is a safe bet that state actors

---

https://github.com/chrisdonahue/sdgan; Phillip Isola et al., Image-to-Image Translation with Conditional Adversarial Nets (Nov. 26, 2018) (unpublished manuscript) (on file with California Law Review); Alec Radford et al., Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks (Jan. 7, 2016) (unpublished manuscript) (on file with California Law Review); Jun-Yan Zhu et al., Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks (Nov. 15, 2018) (unpublished manuscript) (on file with California Law Review).

22.   Aaron van den Oord et al., WaveNet: A Generative Model for Raw Audio (Sept. 19, 2016) (unpublished manuscript) (on file with California Law Review), https://arxiv.org/pdf/1609.03499.pdf [https://perma.cc/QX4W-E6JT].

23.   Ben Popper, *Baidu's New System Can Learn to Imitate Every Accent*, VERGE (Oct. 24, 2017), https://www.theverge.com/2017/10/24/16526370/baidu-deepvoice-3-ai-text-to-speech-voice [https://perma.cc/NXV2-GDVJ].

24.   *See* Chris Donahue et al., Adversarial Audio Synthesis (Feb. 9, 2019) (conference paper), https://arxiv.org/pdf/1802.04208.pdf [https://perma.cc/F5UG-334U]; Yang Gao et al., Voice Impersonation Using Generative Adversarial Networks (Feb. 19, 2018) (unpublished manuscript), https://arxiv.org/abs/1802.06840 [https://perma.cc/5HZV-ZLD3].

25.   *See* Bahar Gholipour, *New AI Tech Can Mimic Any Voice*, SCI. AM. (May 2, 2017), https://www.scientificamerican.com/article/new-ai-tech-can-mimic-any-voice [https://perma.cc/2HSP-83C3]. The ability to cause havoc by using this technology to portray persons saying things they have *never* said looms large. Lyrebird's website includes an "ethics" statement, which defensively invokes notions of technological determinism. The statement argues that impersonation technology is inevitable and that society benefits from gradual introduction to it. *Ethics*, LYREBIRD, https://lyrebird.ai/ethics [https://perma.cc/Q57E-G6MK] ("Imagine that we had decided not to release this technology at all. Others would develop it and who knows if their intentions would be as sincere as ours: they could, for example, only sell the technology to a specific company or an ill-intentioned organization. By contrast, we are making the technology available to anyone and we are introducing it incrementally so that society can adapt to it, leverage its positive aspects for good, while preventing potentially negative applications.").

26.   DARPA's MediFor program is working to "[develop] technologies for the automated assessment of the integrity of an image or video and [integrate] these in an end-to-end media forensics platform." Matt Turek, *Media Forensics (MediFor)*, DEF. ADVANCED RES. PROJECTS AGENCY, https://www.darpa.mil/program/media-forensics [https://perma.cc/VBY5-BQJA]. IARPA's DIVA program is attempting to use artificial intelligence to identify threats by sifting through video imagery. *Deep Intermodal Video Analytics (DIVA) Program*, INTELLIGENCE ADVANCED RES. PROJECTS ACTIVITY, https://www.iarpa.gov/index.php/research-programs/diva [https://perma.cc/4VDX-B68W]. There are no grants from the National Science Foundation awarding federal dollars to researchers studying the detection of doctored audio and video content at this time. E-mail from Seth M. Goldstein, Project Manager, IARPA, Office of the Director of National Intelligence, to Samuel Morse (Apr. 6, 2018, 7:49 AM) (on file with authors).

are conducting classified research in this area. However, it is unclear whether classified research lags behind or outpaces commercial and academic efforts. At the least, we can say with confidence that industry, academia, and governments have the motive, means, and opportunity to push this technology forward at a rapid clip.

B. *Diffusion of Deep-Fake Technology*

The capacity to generate persuasive deep fakes will not stay in the hands of either technologically sophisticated or responsible actors.[27] For better or worse, deep-fake technology will diffuse and democratize rapidly.

As Benjamin Wittes and Gabriella Blum explained in *The Future of Violence: Robots and Germs, Hackers and Drones*, technologies—even dangerous ones—tend to diffuse over time.[28] Firearms developed for state-controlled armed forces are now sold to the public for relatively modest prices.[29] The tendency for technologies to spread only lags if they require scarce inputs that function (or are made to function) as chokepoints to curtail access.[30] Scarcity as a constraint on diffusion works best where the input in question is tangible and hard to obtain; such as plutonium or highly enriched uranium to create nuclear weapons.[31]

Often though, the only scarce input for a new technology is the knowledge behind a novel process or unique data sets. Where the constraint involves an intangible resource like information, preserving secrecy requires not only security against theft, espionage, and mistaken disclosure, but also the capacity and will to keep the information confidential.[32] Depending on the circumstances, the relevant actors may not want to keep the information to themselves and, indeed, may have affirmative commercial or intellectual motivation to disperse it, as in the case of academics or business enterprises.[33]

---

27.   *See* Jaime Dunaway, *Reddit (Finally) Bans Deepfake Communities, but Face-Swapping Porn Isn't Going Anywhere*, SLATE (Feb. 8, 2018), https://slate.com/technology/2018/02/reddit-finally-bans-deepfake-communities-but-face-swapping-porn-isnt-going-anywhere.html [https://perma.cc/A4Z7-2LDF].

28.   *See generally* BENJAMIN WITTES & GABRIELLA BLUM, THE FUTURE OF VIOLENCE: ROBOTS AND GERMS, HACKERS AND DRONES. CONFRONTING A NEW AGE OF THREAT (2015).

29.   *Fresh Air: Assault Style Weapons in the Civilian Market*, NPR (radio broadcast Dec. 20, 2012). Program host Terry Gross interviews Tom Diaz, a policy analyst for the Violence Policy Center. A transcript of the interview can be found at https://www.npr.org/templates/transcript/transcript.php?storyId=167694808 [https://perma.cc/CE3F-5AFX].

30.   *See generally* GRAHAM T. ALLISON ET AL., AVOIDING NUCLEAR ANARCHY (1996).

31.   *Id.*

32.   The techniques that are used to combat cyber attacks and threats are often published in scientific papers, so that a multitude of actors can implement these shields as a defense measure. However, the sophisticated malfeasor can use this information to create cyber weapons that circumvent the defenses that researchers create.

33.   In April 2016, the hacker group "Shadow Brokers" released malware that had allegedly been created by the National Security Agency (NSA). One month later, the malware was used to propagate

Consequently, the capacity to generate deep fakes is sure to diffuse rapidly no matter what efforts are made to safeguard it. The capacity does not depend on scarce tangible inputs, but rather on access to knowledge like GANs and other approaches to machine learning. As the volume and sophistication of publicly available deep-fake research and services increase, user-friendly tools will be developed and propagated online, allowing diffusion to reach beyond experts. Such diffusion has occurred in the past both through commercial and black-market means, as seen with graphic manipulation tools like Photoshop and malware services on the dark web.34 User-friendly capacity to generate deep fakes likely will follow a similar course on both dimensions.35

Indeed, diffusion has begun for deep-fake technology. The recent wave of attention generated by deep fakes began after a Reddit user posted a tool inserting the faces of celebrities into porn videos.36 Once Fake App, "a desktop app for creating photorealistic faceswap videos made with deep learning," appeared online, the public adopted it in short order.37 Following the straightforward steps provided by Fake App, a *New York Times* reporter created a semi-realistic deep-fake video of his face on actor Chris Pratt's body with 1,861 images of himself and 1,023 images of Chris Pratt.38 After enlisting the help of someone with experience blending facial features and source footage, the reporter created a realistic video featuring him as Jimmy Kimmel.39 This portends the diffusion of ever more sophisticated versions of deep-fake technology.

C. *Fueling the Fire*

The capacity to create deep fakes comes at a perilous time. No longer is the public's attention exclusively in the hands of trusted media companies. Individuals peddling deep fakes can quickly reach a massive, even global,

---

the WannaCry cyber attacks, which wreaked havoc on network systems around the globe, threatening to erase files if a ransom was not paid through Bitcoin. *See* Bruce Schneier, *Who Are the Shadow Brokers?*, ATLANTIC (May 23, 2017), https://www.theatlantic.com/technology/archive/2017/05/shadow-brokers/527778 [https://perma.cc/UW2F-V36G].

34.   *See* ARMOR, THE BLACK MARKET REPORT: A LOOK INSIDE THE DARK WEB 2 (2018), https://www.armor.com/app/uploads/2018/03/2018-Q1-Reports-BlackMarket-DIGITAL.pdf [https://perma.cc/4UJA-QJ94] (explaining that the means to conduct a DDoS attack can be purchased for $10/hour, or $200/day).

35.   *See id.*

36.   Emma Grey Ellis, *People Can Put Your Face on Porn—And the Law Can't Help You*, WIRED (Jan. 26, 2018), https://www.wired.com/story/face-swap-porn-legal-limbo [https://perma.cc/B7K7-Y79L].

37.   FAKEAPP, https://www.fakeapp.org.

38.   Kevin Roose, *Here Come the Fake Videos, Too*, N.Y. TIMES (Mar. 4, 2018), https://www.nytimes.com/2018/03/04/technology/fake-videos-deepfakes.html [https://perma.cc/U5QE-EPHX].

39.   *Id.*

audience. As this section explores, networked phenomena, rooted in cognitive bias, will fuel that effort.40

Twenty-five years ago, the practical ability of individuals and organizations to distribute images, audio, and video (whether authentic or not) was limited. In most countries, a handful of media organizations disseminated content on a national or global basis. In the U.S., the major television and radio networks, newspapers, magazines, and book publishers controlled the spread of information.41 While governments, advertisers, and prominent figures could influence mass media, most were left to pursue local distribution of content. For better or worse, relatively few individuals or entities could reach large audiences in this few-to-many information distribution environment.42

The information revolution has disrupted this content distribution model.43 Today, innumerable platforms facilitate global connectivity. Generally speaking, the networked environment blends the few-to-many and many-to-many models of content distribution, democratizing access to communication to an unprecedented degree.44 This reduces the overall amount of gatekeeping, though control still remains with the companies responsible for our digital infrastructure.45 For instance, content platforms have terms-of-service agreements, which ban certain forms of content based on companies' values.46

---

40. *See generally* DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE (2014) [hereinafter CITRON, HATE CRIMES IN CYBERSPACE] (exploring pathologies attendant to online speech including deindividuation, virality, information cascades, group polarization, and filter bubbles). For important early work on filter bubbles, echo chambers, and group polarization in online interactions, see generally ELI PARISER, THE FILTER BUBBLE: WHAT THE INTERNET IS HIDING FROM YOU (2011); CASS R. SUNSTEIN, REPUBLIC.COM (2001).

41. *See generally* NICHOLAS CARR, THE BIG SWITCH: REWIRING THE WORLD, FROM EDISON TO GOOGLE (2008); HOWARD RHEINGOLD, SMART MOBS: THE NEXT SOCIAL REVOLUTION (2002).

42. *See id.*

43. *See generally* SIVA VAIDHYANATHAN, THE GOOGLIZATION OF EVERYTHING (AND WHY WE SHOULD WORRY) (2011).

44. This ably captures the online environment accessible for those living in the United States. As Jack Goldsmith and Tim Wu argued a decade ago, geographic borders and the will of governments can and do make themselves known online. *See generally* JACK GOLDSMITH & TIM WU, WHO OWNS THE INTERNET?: ILLUSIONS OF A BORDERLESS WORLD (2006). The Internet visible in China is vastly different from the Internet visible in the EU, which is different from the Internet visible in the United States (and likely to become more so soon). *See, e.g.,* Elizabeth C. Economy, *The Great Firewall of China: Xi Jinping's Internet Shutdown*, GUARDIAN (June 29, 2018) https://www.theguardian.com/news/2018/jun/29/the-great-firewall-of-china-xi-jinpings-internet-shutdown [https://perma.cc/8GUS-EC59]*;* Casey Newton, *Europe Is Splitting the Internet into Three: How the Copyright Directive Reshapes the Open Web*, VERGE (Mar. 27, 2019, https://www.theverge.com/2019/3/27/18283541/european-union-copyright-directive-Internet-article-13 [https://perma.cc/K235-RZ7Q].

45. Danielle Keats Citron & Neil M. Richards, *Four Principles for Digital Expression (You Won't Believe #3!)*, 95 WASH. U. L. REV. 1353, 1361–64 (2018).

46. *See* CITRON, HATE CRIMES IN CYBERSPACE, *supra* note 40, at 232–35; Danielle Keats Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep*, 93 NOTRE DAME L. REV. 1035, 1037 (2018) [hereinafter Citron, *Extremist Speech*] (noting that platforms' terms of service and community guidelines have banned child pornography, spam, phishing, fraud, impersonation, copyright violations, threats, cyber stalking, nonconsensual pornography, and hate speech); *see also* DANIELLE

They experience pressure from, or adhere to legal mandates of, governments to block or filter certain information like hate speech or "fake news."[47]

Although private companies have enormous power to moderate content (shadow banning it, lowering its prominence, and so on), they may decline to filter or block content that does not amount to obvious illegality. Generally speaking, there is far less screening of content for accuracy, quality, or suppression of facts or opinions that some authority deems undesirable.

Content not only can find its way to online audiences, but can circulate far and wide, sometimes going viral both online and, at times, amplifying further once picked up by traditional media. A variety of cognitive heuristics help fuel these dynamics. Three phenomena in particular—the "information cascade" dynamic, human attraction to negative and novel information, and filter bubbles—help explain why deep fakes may be especially prone to going viral.

First, consider the "information cascade" dynamic.[48] Information cascades result when people stop paying sufficient attention to their own information, relying instead on what they assume others have reliably determined and then passing that information along.. Because people cannot know everything, they often rely on what others say, even if it contradicts their own knowledge.[49] At a certain point, people stop paying attention to their own information and look to what others know.[50] And when people pass along what others think, the

---

KEATS CITRON & QUINTA JURECIC, PLATFORM JUSTICE: CONTENT MODERATION AT AN INFLECTION POINT 12 (Hoover Institution ed., 2018) [hereinafter CITRON & JURECIC, PLATFORM JUSTICE], https://www.hoover.org/sites/default/files/research/docs/citron-jurecic_webreadypdf.pdf [https://perma.cc/M5L6-GNCH] (noting Facebook's Terms of Service agreement banning nonconsensual pornography). *See generally* Danielle Keats Citron, *Cyber Civil Rights*, 89 B.U. L. REV. 61 (2009) [hereinafter Citron, *Cyber Civil Rights*]; Danielle Keats Citron & Helen Norton, *Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age*, 91 B.U. L. REV. 1435, 1458 (2011) (discussing hate speech restrictions contained in platforms' terms of service agreements); Danielle Keats Citron & Benjamin Wittes, *The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity*, 86 FORDHAM L. REV. 401 (2017) (arguing that law should incentivize online platforms to address known illegality in a reasonable manner).

47. *See* Citron, *Extremist Speech*, *supra* note 46, at 1040–49 (exploring pressure from EU Commission on major platforms to remove extremist speech and hate speech). For important work on global censorship efforts, see the scholarship of Anupam Chander, Daphne Keller, and Rebecca McKinnon. *See generally* REBECCA MCKINNON, CONSENT OF THE NETWORKED: THE WORLDWIDE STRUGGLE FOR INTERNET FREEDOM 6 (2012) (arguing that ISPs and online platforms have "far too much power over citizens' lives, in ways that are insufficiently transparent or accountable to the public interest."); Anupam Chander, *Facebookistan*, 90 N.C. L. REV. 1807, 1819–35 (2012); Anupam Chander, *Googling Freedom*, 99 CALIF. L. REV. 1, 5–9 (2011); Daphne Keller, *Toward a Clearer Conversation About Platform Liability*, KNIGHT FIRST AMEND. INST. AT COLUM. U. (April 6, 2018), https://knightcolumbia.org/content/toward-clearer-conversation-about-platform-liability [https://perma.cc/GWM7-J8PW].

48. Carr, *supra* note 41. *See generally* DAVID EASLEY & JON KLEINBERG, NETWORKS, CROWDS, AND MARKETS: REASONING ABOUT A HIGHLY CONNECTED WORLD (2010) (exploring cognitive biases in the information marketplace); CASS SUNSTEIN, REPUBLIC.COM 2.0 (2007) (same).

49. *See generally* EASLEY & KLEINBERG, *supra* note 48.

50. *Id.*

credibility of the original claim snowballs.51 As the cycle repeats, the cascade strengthens.52

Social media platforms are a ripe environment for the formation of information cascades spreading content of all stripes. From there, cascades can spill over to traditional mass-audience outlets that take note of the surge of social media interest and as a result cover a story that otherwise they might not have.53 Social movements have leveraged the power of information cascades, including Black Lives Matter activists54 and the Never Again movement of the Parkland High School students.55 Arab Spring protesters spread videos and photographs of police torture.56 Journalist Howard Rheingold refers to positive information cascades as "smart mobs."57 But not every mob is smart or laudable, and the information cascade dynamic does not account for such distinctions. The Russian covert action program to sow discord in the United States during the 2016 election provides ample demonstration.58

Second, our natural tendency to propagate negative and novel information may enable viral circulation of deep fakes. Negative and novel information "grab[s] our attention as human beings and [] cause[s] us to want to share that information with others—we're attentive to novel threats and especially attentive to negative threats."59 Data scientists, for instance, studied 126,000 news stories shared on Twitter from 2006 to 2010, using third-party fact-checking sites to

---

51. *Id.*

52. *Id.*

53. *See generally* YOCHAI BENKLER, THE WEALTH OF NETWORKS: HOW SOCIAL PRODUCTION TRANSFORMS MARKETS AND FREEDOM (2006) (elaborating the concept of social production in relation to rapid evolution of the information marketplace and resistance to that trend).

54. *See* Monica Anderson & Paul Hitlin, *The Hashtag #BlackLivesMatter Emerges: Social Activism on Twitter*, PEW RES. CTR. (Aug. 15, 2016), http://www.pewInternet.org/2016/08/15/the-hashtag-blacklivesmatter-emerges-social-activism-on-twitter [https://perma.cc/4BW9-L67G] (discussing Black Lives Matter activists' use of the hashtag #BlackLivesMatter to identify their message and display solidarity around race and police use of force).

55. Jonah Engel Bromwich, *How the Parkland Students Got So Good at Social Media*, N.Y. TIMES (Mar. 7, 2018), https://www.nytimes.com/2018/03/07/us/parkland-students-social-media.html [https://perma.cc/7AW9-4HR2] (discussing students' use of social media to keep sustained political attention on the Parkland tragedy).

56. CITRON, HATE CRIMES IN CYBERSPACE, *supra* note 40, at 68.

57. RHEINGOLD, *supra* note 41.

58. The 2018 indictment of the Internet Research Agency in the U.S. District Court for the District of Columbia is available at https://www.justice.gov/file/1035477/download [https://perma.cc/B6WJ-4FLX]; *see also* David A. Graham, *What the Mueller Indictment Reveals*, ATLANTIC (Feb. 16, 2018), https://www.theatlantic.com/politics/archive/2018/02/mueller-roadmap/553604 [https://perma.cc/WU2U-XHWW]; Tim Mak & Audrey McNamara, *Mueller Indictment of Russian Operatives Details Playbook of Information Warfare*, NAT'L PUB. RADIO (Feb. 17, 2018), https://www.npr.org/2018/02/17/586690342/mueller-indictment-of-russian-operatives-details-playbook-of-information-warfare [https://perma.cc/RJ6F-999R].

59. Robinson Meyer, *The Grim Conclusions of the Largest-Ever Study of Fake News*, THE ATLANTIC (Mar. 8, 2018), https://www.theatlantic.com/technology/archive/2018/03/largest-study-ever-fake-news-mit-twitter/555104 [https://perma.cc/PJS2-RKMF].

classify them as true or false.60 According to the study, hoaxes and false rumors reached people ten times faster than accurate stories.61 Even when researchers controlled for differences between accounts originating rumors, falsehoods were 70 percent more likely to get retweeted than accurate news.62 The uneven spread of fake news was not due to bots, which in fact retweeted falsehoods at the same frequency as accurate information.63 Rather, false news spread faster due to *people* retweeting inaccurate news items.64 The study's authors hypothesized that falsehoods had greater traction because they seemed more "novel" and evocative than real news.65 False rumors tended to elicit responses expressing surprise and disgust, while accurate stories evoked replies associated with sadness and trust.66

With human beings seemingly more inclined to spread negative and novel falsehoods, the field is ripe for bots to spur and escalate the spreading of negative misinformation.67 Facebook estimates that as many as 60 million bots may be infesting its platform.68 Bots were responsible for a substantial portion of political content posted during the 2016 election.69 Bots also can manipulate algorithms used to predict potential engagement with content.

Negative information not only is tempting to share, but is also relatively "sticky." As social science research shows, people tend to credit—and remember—negative information far more than positive information.70 Coupled with our natural predisposition towards certain stimuli like sex, gossip, and violence, that tendency provides a welcome environment for harmful deep fakes.71 The Internet amplifies this effect, which helps explain the popularity of

---

60.   Soroush Vosoughi et al., *The Spread of True and False News Online*, 359 SCIENCE 1146, 1146 (2018), http://science.sciencemag.org/content/359/6380/1146/tab-pdf [https://perma.cc/5U5D-UHPZ].

61.   *Id.* at 1148.

62.   *Id.* at 1149.

63.   *Id.* at 1146.

64.   *Id.*

65.   *Id.* at 1149.

66.   *Id.* at 1146, 1150.

67.   Meyer, *supra* note 59 (quoting political scientist Dave Karpf).

68.   Nicholas Confessore et al., *The Follower Factory,* N.Y. TIMES (Jan. 27, 2018), https://www.nytimes.com/interactive/2018/01/27/technology/social-media-bots.html [https://perma.cc/DX34-RENV] ("In November, Facebook disclosed to investors that it had at least twice as many fake users as it previously estimated, indicating that up to 60 million automated accounts may roam the world's largest social media platform."); *see also Extremist Content and Russian Disinformation Online: Working with Tech to Find Solutions: Hearing Before the S. Judiciary Comm.*, 117th   Cong.   (2017)   https://www.judiciary.senate.gov/meetings/extremist-content-and-russian-disinformation-online-working-with-tech-to-find-solutions [https://perma.cc/M5L9-R2MY].

69.   David M. J. Lazer et al., *The Science of Fake News: Addressing Fake News Requires a Multidisciplinary Effort*, 359 SCIENCE 1094, 1095 (2018).

70.   *See, e.g.*, Elizabeth A. Kensinger, *Negative Emotion Enhances Memory Accuracy: Behavioral and Neuroimaging Evidence*, 16 CURRENT DIRECTIONS IN PSYCHOL. SCI. 213, 217 (2007) (finding that "negative emotion conveys focal benefits on memory for detail").

71.   PARISER, *supra* note 40, at 13–14.

gossip sites like TMZ.com.72 Because search engines produce results based on our interests, they tend to feature more of the same—more sex and more gossip.73

Third, filter bubbles further aggravate the spread of false information. Even without the aid of technology, we naturally tend to surround ourselves with information confirming our beliefs. Social media platforms supercharge this tendency by empowering users to endorse and re-share content.74 Platforms' algorithms highlight popular information, especially if it has been shared by friends, and surround us with content from relatively homogenous groups.75 As endorsements and shares accumulate, the chances for an algorithmic boost increase. After seeing friends' recommendations online, individuals tend to pass on those recommendations to their own networks.76 Because people tend to share information with which they agree, social media users are surrounded by information confirming their preexisting beliefs.77 This is what we mean by "filter bubble."78

Filter bubbles can be powerful insulators against the influence of contrary information. In a study of Facebook users, researchers found that individuals reading fact-checking articles had not originally consumed the fake news at issue, and those who consumed fake news in the first place almost never read a fact-check that might debunk it.79

Taken together, common cognitive biases and social media capabilities are behind the viral spread of falsehoods and decay of truth. They have helped entrench what amounts to information tribalism, and the results plague public and private discourse. Information cascades, natural attraction to negative and novel information, and filter bubbles provide an all-too-welcoming environment as deep-fake capacities mature and proliferate.

II.
COSTS AND BENEFITS

Deep-fake technology can and will be used for a wide variety of purposes. Not all will be antisocial; some, in fact, will be profoundly prosocial.

---

72. CITRON, HATE CRIMES IN CYBERSPACE, *supra* note 40, at 68.

73. *Id.*

74. *Id.* at 67.

75. *Id.*

76. *Id.*

77. *Id.*

78. Political scientists Andrew Guess, Brendan Nyhan, and Jason Reifler studied the production and consumption of fake news on Facebook during the 2016 U.S. Presidential election. According to the study, filter bubbles were deep (with one in four individuals visiting from fake news websites), but narrow (the majority of fake news group consumption was concentrated among 10% of the public). *See* ANDREW GUESS ET AL., SELECTIVE EXPOSURE TO MISINFORMATION: EVIDENCE FROM THE CONSUMPTION OF FAKE NEWS DURING THE 2016 U.S. PRESIDENTIAL CAMPAIGN 1 (2018) https://www.dartmouth.edu/~nyhan/fake-news-2016.pdf [https://perma.cc/F3VF-JVCL].

79. *See id.* at 11.

Nevertheless, deep fakes can inflict a remarkable array of harms, many of which are exacerbated by features of the information environment explored above.

A. Beneficial Uses of Deep-Fake Technology

Human ingenuity no doubt will conceive many beneficial uses for deep-fake technology. For now, the most obvious possibilities for beneficial uses fall under the headings of education, art, and the promotion of individual autonomy.

1. Education

Deep-fake technology creates an array of opportunities for educators, including the ability to provide students with information in compelling ways relative to traditional means like readings and lectures. This is similar to an earlier wave of educational innovation made possible by increasing access to ordinary video.[80] With deep fakes, it will be possible to manufacture videos of historical figures speaking directly to students, giving an otherwise unappealing lecture a new lease on life.[81]

Creating modified content will raise interesting questions about intellectual property protections and the reach of the fair use exemption. Setting those obstacles aside, the educational benefits of deep fakes are appealing from a pedagogical perspective in much the same way that is true for the advent of virtual and augmented reality production and viewing technologies.[82]

The technology opens the door to relatively cheap and accessible production of video content that alters existing films or shows, particularly on the audio track, to illustrate a pedagogical point. For example, a scene from a war film could be altered to make it seem that a commander and her legal advisor are discussing application of the laws of war, when in the original the dialogue had nothing to do with that—and the scene could be re-run again and again with modifications to the dialogue tracking changes to the hypothetical scenario under

---

80.   Emily Cruse, Using Educational Video in the Classroom: Theory, Research, and Practice, 1-2 (2013) (unpublished manuscript), https://www.safarimontage.com/pdfs/training/UsingEducationalVideoInTheClassroom.pdf [https://perma.cc/AJ8Q-WZP4].

81.   Face2Face is a real-time face capture and reenactment software developed by researchers at the University of Erlangen-Nuremberg, the Max-Planck-Institute for Informatics, and Stanford University. The applications of this technology could reinvent the way students learn about historical events and figures. *See* Justus Thies et al., Face2Face: Real-time Face Capture and Reenactment of RGB Videos (June 2016) (29th IEEE-CVPR 2016 conference paper), http://www.graphics.stanford.edu/~niessner/papers/2016/1facetoface/thies2016face.pdf [https://perma.cc/S94K-DPU5].

82.   Adam Evans, *Pros and Cons of Virtual Reality in the Classroom,* CHRON. HIGHER EDUC. (Apr. 8, 2018), https://www.chronicle.com/article/ProsCons-of-Virtual/243016 [https://perma.cc/TN84-89SQ].

CALIFORNIA LAW REVIEW [Vol. 107:1753

consideration. If done well, it would surely beat just having the professor asking students to imagine the shifting scenario out of whole cloth.[83]

The educational value of deep fakes will extend beyond the classroom. In the spring of 2018, Buzzfeed provided an apt example when it circulated a video that appeared to feature Barack Obama warning of the dangers of deep-fake technology itself.[84] One can imagine deep fakes deployed to support educational campaigns by public-interest organizations such as Mothers Against Drunk Driving.

2. Art

The potential artistic benefits of deep-fake technology relate to its educational benefits, though they need not serve any formal educational purpose. Thanks to the use of existing technologies that resurrect dead performers for fresh roles, the benefits to creativity are already familiar to mass audiences.[85] For example, the startling appearance of the long-dead Peter Cushing as the venerable Grand Moff Tarkin in 2016's *Rogue One* was made possible by a deft combination of live acting and technical wizardry. That prominent illustration delighted some and upset others.[86] The *Star Wars* contribution to this theme continued in *The Last Jedi* when Carrie Fisher's death led the filmmakers to fake additional dialogue using snippets from real recordings.[87]

Not all artistic uses of deep-fake technologies will have commercial potential. Artists may find it appealing to express ideas through deep fakes, including, but not limited to, productions showing incongruities between apparent speakers and their apparent speech. Video artists might use deep-fake technology to satirize, parody, and critique public figures and public officials. Activists could use deep fakes to demonstrate their point in a way that words alone could not.

3. Autonomy

Just as art overlaps with education, deep fakes implicate self-expression. But not all uses of deep fakes for self-expression are best understood as art. Some

---

83. The facial animation software CrazyTalk, by Reallusion, animates faces from photographs or cartoons and can be used by educators to further pedagogical goals. The software is available at https://www.reallusion.com/crazytalk/default.html [https://perma.cc/TTX8-QMJP].

84. *See* Choi, *supra* note 16.

85. Indeed, film contracts now increasingly address future uses of a person's image in subsequent films via deep fake technology in the event of their death.

86. Dave Itzkoff, *How 'Rogue One' Brought Back Familiar Faces*, N.Y. TIMES (Dec. 27, 2016), https://www.nytimes.com/2016/12/27/movies/how-rogue-one-brought-back-grand-moff-tarkin.html [https://perma.cc/F53C-TDYV].

87. Evan Narcisse, *It Took Some Movie Magic to Complete Carrie Fisher's Leia Dialogue in* The Last Jedi, GIZMODO (Dec. 8, 2017), https://io9.gizmodo.com/it-took-some-movie-magic-to-complete-carrie-fishers-lei-1821121635 [https://perma.cc/NF5H-GPJF].

may be used to facilitate "avatar" experiences for a variety of self-expressive ends that might best be described in terms of autonomy.

Perhaps most notably, deep-fake audio technology holds promise to restore the ability of persons suffering from certain forms of paralysis, such as ALS, to speak with their own voice.[88] Separately, individuals suffering from certain physical disabilities might interpose their faces and that of consenting partners into pornographic videos, enabling virtual engagement with an aspect of life unavailable to them in a conventional sense.[89]

The utility of deep-fake technology for avatar experiences, which need not be limited to sex, closely relates to more familiar examples of technology. Video games, for example, enable a person to have or perceive experiences that might otherwise be impossible, dangerous, or otherwise undesirable if pursued in person. The customizable avatars from Nintendo Wii (known as "Mii") provide a familiar and non-threatening example. The video game example underscores that the avatar scenario is not always a serious matter, and sometimes boils down to no more and no less than the pursuit of happiness.

Deep-fake technology confers the ability to integrate more realistic simulacrums of one's own self into an array of media, thus producing a stronger avatar effect. For some aspects of the pursuit of autonomy, this will be a very good thing (as the book and film *Ready Player One* suggests, albeit with reference to a vision of advanced virtual reality rather than deep-fake technology). Not so for others, however. Indeed, as we describe below, the prospects for the harmful use of deep-fake technology are legion.

B. Harmful Uses of Deep-Fake Technology

Human ingenuity, alas, is not limited to applying technology to beneficial ends. Like any technology, deep fakes also will be used to cause a broad spectrum of serious harms, many of them exacerbated by the combination of networked information systems and cognitive biases described above.

1. Harm to Individuals or Organizations

Lies about what other people have said or done are as old as human society, and come in many shapes and sizes. Some merely irritate or embarrass, while others humiliate and destroy; some spur violence. All of this will be true with deep fakes as well, only more so due to their inherent credibility and the manner

---

88.   Sima Shakeri, *Lyrebird Helps ALS Ice Bucket Challenge Co-Founder Pat Quinn Get His Voice Back: Project Revoice Can Change Lives*, HUFFINGTON POST (Apr. 14, 2018), https://www.huffingtonpost.ca/2018/04/14/lyrebird-helps-als-ice-bucket-challenge-co-founder-pat-quinn-get-his-voice-back_a_23411403 [https://perma.cc/R5SD-Y37Y].

89.   *See* Allie Volpe, *Deepfake Porn has Terrifying Implications. But What if it Could Be Used for Good?*, MEN'S HEALTH (Apr. 13, 2018), https://www.menshealth.com/sex-women/a19755663/deepfakes-porn-reddit-pornhub [https://perma.cc/EFX9-2BUE].

in which they hide the liar's creative role. Deep fakes will emerge as powerful mechanisms for some to exploit and sabotage others.

a. *Exploitation*

There will be no shortage of harmful exploitations. Some will be in the nature of theft, such as stealing people's identities to extract financial or some other benefit. Others will be in the nature of abuse, commandeering a person's identity to harm them or individuals who care about them. And some will involve both dimensions, whether the person creating the fake so intended or not.

As an example of extracting value, consider the possibilities for the realm of extortion. Blackmailers might use deep fakes to extract something of value from people, even those who might normally have little or nothing to fear in this regard, who (quite reasonably) doubt their ability to debunk the fakes persuasively, or who fear that any debunking would fail to reach far and fast enough to prevent or undo the initial damage.[90] In that case, victims might be forced to provide money, business secrets, or nude images or videos (a practice known as sextortion) to prevent the release of the deep fakes.[91] Likewise, fraudulent kidnapping claims might prove more effective in extracting ransom when backed by video or audio appearing to depict a victim who is not in fact in the fraudster's control.

Not all value extraction takes a tangible form. Deep-fake technology can also be used to exploit an individual's sexual identity for other's gratification.[92] Thanks to deep-fake technology, an individual's face, voice, and body can be swapped into real pornography.[93] A subreddit (now closed) featured deep-fake sex videos of female celebrities and amassed more than 100,000 users.[94] As one Reddit user asked, "I want to make a porn video with my ex-girlfriend. But I

---

90.   *See generally* ADAM DODGE & ERICA JOHNSTONE, USING FAKE VIDEO TECHNOLOGY TO PERPETUATE INTIMATE PARTNER ABUSE 6 (2018), http://withoutmyconsent.org/blog/new-advisory-helps-domestic-violence-survivors-prevent-and-stop-deepfake-abuse [https://perma.cc/K3Y2-XG2Q] (discussing how deep fakes used as black mail of an intimate partner could violate the California Family Code). The advisory was published by the non-profit organization Without My Consent, which combats online invasions of privacy.

91.   Sextortion thrives on the threat that the extortionist will disclose sex videos or nude images unless more nude images or videos are provided. BENAJMIN WITTES ET AL., SEXTORTION: CYBERSECURITY, TEENAGERS, AND REMOTE SEXUAL ASSAULT (Brookings Inst. ed., 2016), https://www.brookings.edu/wp-content/uploads/2016/05/sextortion1-1.pdf [https://perma.cc/7K9N-5W7C].

92.   *See* DODGE & JOHNSTONE, *supra* note 90, at 6 (explaining the likelihood that domestic abusers and cyber stalkers will use deep sex tapes to harm victims); Janko Roettgers, '*Deep Fakes' Will Create Hollywood's Next Sex Tape Scare*, VARIETY (Feb. 2, 2018), http://variety.com/2018/digital/news/hollywood-sex-tapes-deepfakes-ai-1202685655 [https://perma.cc/98HQ-668G].

93.   Danielle Keats Citron, *Sexual Privacy*, 128 YALE L. J. 1870, 1921–24 (2019) [hereinafter Citron, *Sexual Privacy*].

94.   DODGE & JOHNSTONE, *supra* note 90, at 6.

don't have any high-quality video with her, but I have lots of good photos."[95] A Discord user explained that he made a "pretty good" video of a girl he went to high school with, using around 380 photos scraped from her Instagram and Facebook accounts.[96]

These examples highlight an important point: the gendered dimension of the exploitation of deep fakes. In all likelihood, the majority of victims of fake sex videos will be female. This has been the case for cyber stalking and non-consensual pornography, and likely will be the case for deep-fake sex videos.[97]

One can easily imagine deep-fake sex videos subjecting individuals to violent, humiliating sex acts. This shows that not all such fakes will be designed primarily, or at all, for the creator's sexual or financial gratification. Some will be nothing less than cruel weapons meant to terrorize and inflict pain. Of deep-fake sex videos, Mary Anne Franks has astutely said, "If you were the worst misogynist in the world, this technology would allow you to accomplish whatever you wanted."[98]

When victims discover that they have been used in deep-fake sex videos, the psychological damage may be profound—whether or not this was the video creator's aim. Victims may feel humiliated and scared.[99] Deep-fake sex videos force individuals into virtual sex, reducing them to sex objects. As Robin West has observed, threats of sexual violence "literally, albeit not physically, penetrates the body."[100] Deep-fake sex videos can transform rape threats into a terrifying virtual reality. They send the message that victims can be sexually abused at whim. Given the stigma of nude images, especially for women and girls, individuals depicted in fake sex videos also may suffer collateral consequences in the job market, among other places, as we explain in more detail below in our discussion of sabotage.[101]

---

95. *Id.*

96. *Id.*

97. ASIA A. EATON ET AL., 2017 NATIONWIDE ONLINE STUDY OF NONCONSENSUAL PORN VICTIMIZATION AND PERPETRATION 12 (Cyber C.R. Initiative ed., 2017), https://www.cybercivilrights.org/wp-content/uploads/2017/06/CCRI-2017-Research-Report.pdf [https://perma.cc/2HYP-7ELV] ("Women were significantly more likely [1.7 times] to have been victims of [non-consensual porn] or to have been threatened with [non-consensual porn]. . . .").

98. Drew Harwell, *Fake-Porn Videos Are Being Weaponized to Harass and Humiliate Women: 'Everybody is a Potential Target'*, WASH. POST (Dec. 30, 2018), https://www.washingtonpost.com/technology/2018/12/30/fake-porn-videos-are-being-weaponized-harass-humiliate-women-everybody-is-potential-target/?utm_term=.936bfc339777 [https://perma.cc/D37Y-DPXB].

99. *See generally* Rana Ayyub, *In India, Journalists Face Slut-Shaming and Rape Threats*, N.Y. TIMES (May 22, 2018), https://www.nytimes.com/2018/05/22/opinion/india-journalists-slut-shaming-rape.html [https://perma.cc/A7WR-PF6L]; *'I Couldn't Talk or Sleep for Three Days': Journalist Rana Ayyub's Horrific Social Media Ordeal over Fake Tweet*, Daily O (Apr. 26, 2018), https://www.dailyo.in/variety/rana-ayyub-trolling-fake-tweet-social-media-harassment-hindutva/story/1/23733.html [https://perma.cc/J6G6-H6GZ].

100. ROBIN WEST, CARING FOR JUSTICE 102–03 (1997) (emphasis omitted).

101. Deep-fake sex videos should be considered in light of the broader cyber stalking phenomenon, which more often targets women and usually involves online assaults that are sexually

These examples are but the tip of a disturbing iceberg. Like sexualized deep fakes, imagery depicting non-sexual abuse or violence might also be used to threaten, intimidate, and inflict psychological harm on the depicted victim (or those who care for that person). Deep fakes also might be used to portray someone, falsely, as endorsing a product, service, idea, or politician. Other forms of exploitation will abound.

b.   Sabotage

In addition to inflicting direct psychological harm on victims, deep-fake technology can be used to harm victims along other dimensions due to their utility for reputational sabotage. Across every field of competition—workplace, romance, sports, marketplace, and politics—people will have the capacity to deal significant blows to the prospects of their rivals.

It could mean the loss of romantic opportunity, the support of friends, the denial of a promotion, the cancellation of a business opportunity, and beyond. Deep-fake videos could depict a person destroying property in a drunken rage. They could show people stealing from a store; yelling vile, racist epithets; using drugs; or any manner of antisocial or embarrassing behavior like sounding incoherent. Depending on the circumstances, timing, and circulation of the fake, the effects could be devastating.

In some instances, debunking the fake may come too late to remedy the initial harm. For example, consider how a rival might torpedo the draft position of a top pro sports prospect by releasing a compromising deep-fake video just as the draft begins. Even if the video is later doubted as a fake, it could be impossible to undo the consequences (which might involve the loss of millions of dollars) because once cautious teams make other picks, the victim may fall into later rounds of the draft (or out of the draft altogether).[102]

The nature of today's communication environment enhances the capacity of deep fakes to cause reputational harm. The combination of cognitive biases and algorithmic boosting increases the chances for salacious fakes to circulate. The ease of copying and storing data online—including storage in remote jurisdictions—makes it much harder to eliminate fakes once they are posted and shared. These considerations combined with the ever-improving search engines increase the chances that employers, business partners, or romantic interests will encounter the fake.

---

threatening and sexually demeaning. *See* CITRON, HATE CRIMES IN CYBERSPACE, *supra* note 40, at 13–19.

    102.   This hypothetical is modeled on an actual event, albeit one involving a genuine rather than a falsified compromising video. In 2016, a highly regarded NFL prospect named Laremy Tunsill may have lost as much as $16 million when, on the verge of the NFL draft, someone released a video showing him smoking marijuana with a bong and gas mask. *See* Jack Holmes, *A Hacker's Tweet May Have Cost This NFL Prospect Almost $16 Million*, ESQUIRE (Apr. 29, 2016), https://www.esquire.com/sports/news/a44457/laremy-tunsil-nfl-draft-weed-lost-millions [https://perma.cc/7PEL-PRBF].

Once discovered, deep fakes can be devastating to those searching for employment. Search results matter to employers.103 According to a 2009 Microsoft study, more than 90 percent of employers use search results to make decisions about candidates, and in more than 77 percent of cases, those results have a negative result. As the study explained, employers often decline to interview or hire people because their search results featured "inappropriate photos."104 The reason for those results should be obvious. It is less risky and expensive to hire people who do not have the baggage of damaged online reputations. This is especially true in fields where the competition for jobs is steep.105 There is little reason to think the dynamics would be significantly different with respect to romantic prospects.106

Deep fakes can be used to sabotage business competitors. Deep-fake videos could show a rival company's chief executive engaged in any manner of disreputable behavior, from purchasing illegal drugs to hiring underage prostitutes to uttering racial epithets to bribing government officials. Deep fakes could be released just in time to interfere with merger discussions or bids for government contracts. As with the sports draft example, mundane business opportunities could be thwarted even if the videos are ultimately exposed as fakes.

---

103.   *Number of Employers Using Social Media to Screen Candidates at All-Time High, Finds Latest CareerBuilder Study*, CAREERBUILDER: PRESS ROOM (June 15, 2017), http://press.careerbuilder.com/2017-06-15-Number-of-Employers-Using-Social-Media-to-Screen-Candidates-at-All-Time-High-Finds-Latest-CareerBuilder-Study [https://perma.cc/K6BD-DYSV] (noting that a national survey conducted in 2017 found that over half of employers will not hire a candidate without an online presence and may choose not to hire a candidate based on negative social media content).

104.   This has been the case for nude photos posted without consent, often known as revenge porn. *See generally* CITRON, HATE CRIMES IN CYBERSPACE, *supra* note 40, at 17–18, 48–49 (exploring the economic fallout of the nonconsensual posting of someone's nude image); Mary Anne Franks, *"Revenge Porn" Reform: A View from the Front Lines*, 69 FLA. L. REV. 1251, 1308–23 (2017). For recent examples, see Tasneem Nashrulla, *A Middle School Teacher Was Fired After a Student Obtained Her Topless Selfie. Now She is Suing the School District for Gender Discrimination*, BUZZFEED (Apr. 4, 2019), https://www.buzzfeednews.com/article/tasneemnashrulla/middle-school-teacher-fired-topless-selfie-lawsuit [https://perma.cc/3PGZ-CZ5R]; Annie Seifullah, *Revenge Porn Took My Career. The Law Couldn't Get It Back*, JEZEBEL (July 18, 2018), https://jezebel.com/revenge-porn-took-my-career-the-law-couldnt-get-it-bac-1827572768 [https://perma.cc/D9Y8-63WH].

105.   *See* Danielle Keats Citron & Mary Anne Franks, *Criminalizing Revenge Porn*, 49 WAKE FOREST L. REV. 345, 352–53 (2014) ("Most employers rely on candidates' online reputations as an employment screen.").

106.   Journalist Rana Ayuub, who faced vicious online abuse including her image in deep-fake sex videos, explained that the deep fakes seemed designed to label her as "promiscuous," "immoral," and damaged goods. Ayyub, *supra* note 99. *See generally* Citron, *Sexual Privacy*, supra note 93, at 1925–26 (discussing how victims of deep-fake sex videos felt crippled and unable to talk or eat, let alone engage with others); Danielle Keats Citron, *Why Sexual Privacy Matters for Trust*, WASH. U. L. REV. (forthcoming) (recounting fear of dating and embarrassment experienced by individuals whose nude photos were disclosed online without consent).

2. *Harm to Society*

Deep fakes are not just a threat to specific individuals or entities. They have the capacity to harm society in a variety of ways. Consider the following:

- Fake videos could feature public officials taking bribes, displaying racism, or engaging in adultery.

- Politicians and other government officials could appear in locations where they were not, saying or doing things that they did not.[107]

- Fake audio or video could involve damaging campaign material that claims to emanate from a political candidate when it does not.[108]

- Fake videos could place them in meetings with spies or criminals, launching public outrage, criminal investigations, or both.

- Soldiers could be shown murdering innocent civilians in a war zone, precipitating waves of violence and even strategic harms to a war effort.[109]

- A deep fake might falsely depict a white police officer shooting an unarmed black man while shouting racial epithets.

- A fake audio clip might "reveal" criminal behavior by a candidate on the eve of an election.

- Falsified video appearing to show a Muslim man at a local mosque celebrating the Islamic State could stoke distrust of, or even violence against, that community.

- A fake video might portray an Israeli official doing or saying something so inflammatory as to cause riots in neighboring countries, potentially disrupting diplomatic ties or sparking a wave of violence.

- False audio might convincingly depict U.S. officials privately "admitting" a plan to commit an outrage overseas, timed to disrupt an important diplomatic initiative.

- A fake video might depict emergency officials "announcing" an impending missile strike on Los Angeles or an emergent pandemic in New York City, provoking panic and worse.

---

107. *See, e.g.*, Linton Weeks, *A Very Weird Photo of Ulysses S. Grant,* NAT'L PUB. RADIO (Oct. 27, 2015 11:03 AM), https://www.npr.org/sections/npr-history-dept/2015/10/27/452089384/a-very-weird-photo-of-ulysses-s-grant [https://perma.cc/F3U6-WRVF] (discussing a doctored photo of Ulysses S. Grant from the Library of Congress archives that was created over 100 years ago).

108. For powerful work on the potential damage of deep-fake campaign speech, see Rebecca Green, *Counterfeit Campaign Speech*, 70 HASTINGS L. J. (forthcoming 2019).

109. *Cf.* Vindu Goel and Sheera Frenkel, *In India Election, False Posts and Hate Speech Flummox Facebook*, N.Y TIMES (Apr. 1, 2019), https://www.nytimes.com/2019/04/01/technology/india-elections-facebook.html [https://perma.cc/55AW-X6Q3].

As these scenarios suggest, the threats posed by deep fakes have systemic dimensions. The damage may extend to, among other things, distortion of democratic discourse on important policy questions; manipulation of elections; erosion of trust in significant public and private institutions; enhancement and exploitation of social divisions; harm to specific military or intelligence operations or capabilities; threats to the economy; and damage to international relations.

a.   *Distortion of Democratic Discourse*

Public discourse on questions of policy currently suffers from the circulation of false information.110 Sometimes lies are intended to undermine the credibility of participants in such debates, and sometimes lies erode the factual foundation that ought to inform policy discourse. Even without prevalent deep fakes, information pathologies abound. But deep fakes will exacerbate matters by raising the stakes for the "fake news" phenomenon in dramatic fashion (quite literally).111

Many actors will have sufficient interest to exploit the capacity of deep fakes to skew information and thus manipulate beliefs. As recent actions by the Russian government demonstrate, state actors sometimes have such interests.112 Other actors will do it as a form of unfair competition in the battle of ideas. And others will do it simply as a tactic of intellectual vandalism and fraud. The combined effects may be significant, including but not limited to the disruption of elections. But elections are vulnerable to deep fakes in a separate and distinctive way as well, as we will explore in the next section.

Democratic discourse is most functional when debates build from a foundation of shared facts and truths supported by empirical evidence.113 In the absence of an agreed upon reality, efforts to solve national and global problems become enmeshed in needless first-order questions like whether climate change is real.114 The large-scale erosion of public faith in data and statistics has led us

---

110.   *See* Steve Lohr, *It's True: False News Spreads Faster and Wider. And Humans Are to Blame*, N.Y. TIMES (Mar. 8, 2018), https://www.nytimes.com/2018/03/08/technology/twitter-fake-news-research.html [https://perma.cc/AB74-CUWV].

111.   Franklin Foer, *The Era of Fake Video Begins*, ATLANTIC (May 2018), https://www.theatlantic.com/magazine/archive/2018/05/realitys-end/556877 [https://perma.cc/RX2A-X8EE] ("Fabricated videos will create new and understandable suspicions about everything we watch. Politicians and publicists will exploit those doubts. When captured in a moment of wrongdoing, a culprit will simply declare the visual evidence a malicious concoction.").

112.   Charlie Warzel, *2017 Was the Year Our Internet Destroyed Our Shared Reality,* BUZZFEED (Dec. 28, 2017), https://www.buzzfeed.com/charliewarzel/2017-year-the-Internet-destroyed-shared-reality?utm_term=.nebaDjYmj [https://perma.cc/8WWS-UC8K].

113.   Mark Verstraete & Derek E. Bambauer, *Ecosystem of Distrust*, 16 FIRST AMEND. L. REV. 129, 152 (2017). For powerful scholarship on how lies undermine culture of trust, see SEANA VALENTINE SHRIFFIN, SPEECH MATTERS: ON LYING, MORALITY, AND THE LAW (2014).

114.   Verstraete & Bambauer, *supra* note 113, at 144 ("Trust in data and statistics is a precondition to being able to resolve disputes about the world—they allow participants in policy debates to operate at least from a shared reality.").

to a point where the simple introduction of empirical evidence can alienate those who have come to view statistics as elitist.[115] Deep fakes will allow individuals to live in their own subjective realities, where beliefs can be supported by manufactured "facts." When basic empirical insights provoke heated contestation, democratic discourse has difficulty proceeding. In a marketplace of ideas flooded with deep-fake videos and audio, truthful facts will have difficulty emerging from the scrum.

b.   *Manipulation of Elections*

In addition to the ability of deep fakes to inject visual and audio falsehoods into policy debates, a deeply convincing variation of a long-standing problem in politics, deep fakes can enable a particularly disturbing form of sabotage: distribution of a damaging, but false, video or audio about a political candidate. The potential to sway the outcome of an election is real, particularly if the attacker is able to time the distribution such that there will be enough window for the fake to circulate but not enough window for the victim to debunk it effectively (assuming it can be debunked at all). In this respect, the election scenario is akin to the NBA draft scenario described earlier. Both involve decisional chokepoints: narrow windows of time during which irrevocable decisions are made, and during which the circulation of false information therefore may have irremediable effects.

The 2017 election in France illustrates the perils. In this variant of the operation executed against the Clinton campaign in the United States in 2016, the Russians mounted a covert-action program that blended cyber-espionage and information manipulation in an effort to prevent the election of Emmanuel Macron as President of France in 2017.[116] The campaign included theft of large numbers of digital communications and documents, alteration of some of those documents in hopes of making them seem problematic, and dumping a lot of them on the public alongside aggressive spin. The effort ultimately fizzled for many reasons, including: poor tradecraft that made it easy to trace the attack; smart defensive work by the Macron team, which planted their own false documents throughout their own system to create a smokescreen of distrust; a lack of sufficiently provocative material despite an effort by the Russians to engineer scandal by altering some of the documents prior to release; and mismanagement of the timing of the document dump, which left enough time for the Macron team and the media to discover and point out all these flaws.[117]

---

115.   *Id.*

116.   *See* Aurelien Breeden et al., *Macron Campaign Says It Was Target of 'Massive' Hacking Attack*, N.Y. TIMES (May 5, 2017), https://www.nytimes.com/2017/05/05/world/europe/france-macron-hacking.html [https://perma.cc/4RC8-PV5G].

117.   *See, e.g.*, Adam Nossiter et al., *Hackers Came, But the French Were Prepared,* N.Y. TIMES (May 9, 2017), https://www.nytimes.com/2017/05/09/world/europe/hackers-came-but-the-french-were-prepared.html [https://perma.cc/P3EW-H5ZY].

It was a bullet dodged, yes, but a bullet nonetheless. The Russians could have acted with greater care, both in terms of timing and tradecraft. They could have produced a more-damning fake document, for example, dropping it just as polls opened. Worse, they could have distributed a deep fake consisting of seemingly-real video or audio evidence persuasively depicting Macron speaking or doing something shocking.

This version of the deep-fake threat is not limited to state-sponsored covert action. States may have a strong incentive to develop and deploy such tools to sway elections, but there will be no shortage of non-state actors and individuals motivated to do the same. The limitation on such interventions has much more to do with means than motive, as things currently stand. The diffusion of the capacity to produce high-quality deep fakes will erode that limitation, empowering an ever-widening circle of participants to inject false-but-compelling information into a ready and willing information-sharing environment. If executed and timed well enough, such interventions are bound to tip an outcome sooner or later—and in a larger set of cases they will at least cast a shadow of illegitimacy over the election process itself.

c.   Eroding Trust in Institutions

Deep fakes will erode trust in a wide range of both public and private institutions and such trust will become harder to maintain. The list of public institutions for which this will matter runs the gamut, including elected officials, appointed officials, judges, juries, legislators, staffers, and agencies. One can readily imagine, in the current climate especially, a fake-but-viral video purporting to show FBI special agents discussing ways to abuse their authority to pursue a Trump family member. Conversely, we might see a fraudulent video of ICE officers speaking with racist language about immigrants or acting cruelly towards a detained child. Particularly where strong narratives of distrust already exist, provocative deep fakes will find a primed audience.

Private sector institutions will be just as vulnerable. If an institution has a significant voice or role in society, whether nationally or locally, it is a potential target. More to the point, such institutions already are subject to reputational attacks, but soon will have to face abuse in the form of deep fakes that are harder to debunk and more likely to circulate widely. Religious institutions are an obvious target, as are politically-engaged entities ranging from Planned Parenthood to the NRA.[118]

---

118.   Recall that the Center for Medical Progress released videos of Planned Parenthood officials that Planned Parenthood argued had been deceptively edited to embarrass the organization. *See, e.g.*, Jackie Calmes, *Planned Parenthood Videos Were Altered, Analysis Finds*, N.Y. TIMES (Aug. 27, 2015), https://www.nytimes.com/2015/08/28/us/abortion-planned-parenthood-videos.html [https://perma.cc/G52X-V8ND]. Imagine the potential for deep fakes designed for such a purpose.

d. *Exacerbating Social Divisions*

The institutional examples relate closely to significant cleavages in American society involving identity and policy commitments. Indeed, this is what makes institutions attractive targets for falsehoods. As divisions become entrenched, the likelihood that opponents will believe negative things about the other side—and that some will be willing to spread lies towards that end—grows.[119] However, institutions will not be the only ones targeted with deep fakes. We anticipate that deep fakes will reinforce and exacerbate the underlying social divisions that fueled them in the first place.

Some have argued that this was the actual—or at least the original—goal of the Russian covert action program involving intervention in American politics in 2016. The Russians may have intended to enhance American social divisions as a general proposition, rendering us less capable of forming consensus on important policy questions and thus more distracted by internal squabbles.[120] Texas is illustrative.[121] Russia promoted conspiracy theories about federal military power during the innocuous, "Jade Helm" training exercises.[122] Russian operators organized an event in Houston to protest radical Islam and a counter-protest of that event;[123] they also promoted a Texas independence movement.[124] Deep fakes will strengthen the hand of those who seek to divide us in this way.

Deep fakes will not merely add fuel to the fire sustaining divisions. In some instances, the emotional punch of a fake video or audio might accomplish a degree of mobilization-to-action that written words alone could not.[125] Consider

---

119. *See* Brian E. Weeks, *Emotions, Partisanship, and Misperceptions: How Anger and Anxiety Moderate the Effect of Partisan Bias on Susceptibility to Political Misinformation*, 65 J. COMM. 699, 711–15 (2015) (discussing how political actors can spread political misinformation by recognizing and exploiting common human emotional states).

120. JON WHITE, DISMISS, DISTORT, DISTRACT, AND DISMAY: CONTINUITY AND CHANGE IN RUSSIAN DISINFORMATION (Inst. for European Studies ed. 2016), https://www.ies.be/node/3689 [https://perma.cc/P889-768J].

121. The CalExit campaign is another illustration of Russian disinformation campaign. *'Russian Trolls' Promoted California Independence*, BBC (Nov. 4, 2017), http://www.bbc.com/news/blogs-trending-41853131 [https://perma.cc/68Q8-KNDG].

122. Cassandra Pollock & Alex Samuels, *Hysteria Over Jade Helm Exercise in Texas Was Fueled by Russians, Former CIA Director Says*, TEX. TRIB. (May 3, 2018), https://www.texastribune.org/2018/05/03/hysteria-over-jade-helm-exercise-texas-was-fueled-russians-former-cia [https://perma.cc/BU2Y-E7EY].

123. Scott Shane, *How Unwitting Americans Encountered Russian Operatives Online*, N.Y. TIMES (Feb. 18, 2018), https://www.nytimes.com/2018/02/18/us/politics/russian-operatives-facebook-twitter.html [https://perma.cc/4C8Y-STP7].

124. Casey Michel, *How the Russians Pretended to Be Texans—And Texans Believed Them*, WASH. POST (Oct. 17, 2017), https://www.washingtonpost.com/news/democracy-post/wp/2017/10/17/how-the-russians-pretended-to-be-texans-and-texans-believed-them/?noredirect=on&utm_term=.4730a395a684 [https://perma.cc/3Q7V-8YZK].

125. The "Pizzagate" conspiracy theory is a perfect example. There, an individual stormed a D.C. restaurant with a gun because online stories falsely claimed that Presidential candidate Hillary Clinton ran a child sex exploitation ring out of its basement. *See* Marc Fisher et al., *Pizzagate: From Rumor, to Hashtag, to Gunfire in D.C.*, WASH. POST (Dec. 6, 2016),

a situation of fraught, race-related tensions involving a police force and a local community. A sufficiently inflammatory deep fake depicting a police officer using racial slurs, shooting an unarmed person, or both could set off substantial civil unrest, riots, or worse. Of course, the same deep fake might be done in reverse, falsely depicting a community leader calling for violence against the police. Such events would impose intangible costs by sharpening societal divisions, as well as tangible costs for those tricked into certain actions and those suffering from those actions.

e.   *Undermining Public Safety*

The foregoing example illustrates how a deep fake might be used to enhance social divisions and to spark actions—even violence—that fray our social fabric. But note, too, how deep fakes can undermine public safety.

A century ago, Justice Oliver Wendell Holmes warned of the danger of falsely shouting fire in a crowded theater.[126] Now, false cries in the form of deep fakes go viral, fueled by the persuasive power of hyper-realistic evidence in conjunction with the distribution powers of social media.[127] The panic and damage Holmes imagined may be modest in comparison to the potential unrest and destruction created by a well-timed deep fake.[128]

In the best-case scenario, real public panic might simply entail economic harms and hassles. In the worst-case scenario, it might involve property destruction, personal injuries, and/or death. Deep fakes increase the chances that someone can induce a public panic.

They might not even need to capitalize on social divisions to do so. In early 2018, we saw a glimpse of how a panic might be caused through ordinary human error when an employee of Hawaii's Emergency Management Agency issued a

---

https://www.washingtonpost.com/local/pizzagate-from-rumor-to-hashtag-to-gunfire-in-dc/2016/12/06/4c7def50-bbd4-11e6-94ac-3d324840106c_story.html [https://perma.cc/FV7W-PC9W].

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

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

128.   In our keynote at the University of Maryland Law Review symposium inspired by this article, we brought the issue close to home (for one of us) in Baltimore—the death of Freddie Gray while he was in policy custody. We asked the audience: "Imagine if a deep-fake video appeared of the police officers responsible for Mr. Gray's death in which they said they were ordered to kill Mr. Gray. As most readers know, the day after Mr. Gray's death was characterized by protests and civil unrest. If such a deep-fake video had appeared and gone viral, we might have seen far more violence and disruption in Baltimore. If the timing was just right and the video sufficiently inflammatory, we might have seen greater destruction of property and possibly of lives." Robert Chesney & Danielle Keats Citron, *21st Century Style Truth Decay: Deep Fakes and the Challenge for Privacy, Free Expression, and National Security*, 78 MD. L. REV. 887 (2019); *see also* Maryland Carey Law, *Truth Decay– Maryland Law Review Keynote Symposium Address*, YOUTUBE (Feb. 6, 2019), https://www.youtube.com/watch?v=WrYlKHiWv2c [https://perma.cc/T28M-ZBBN].

warning to the public about an incoming ballistic missile.129 Less widely noted, we saw purposeful attempts to induce panic when the Russian Internet Research Agency mounted a sophisticated and well-resourced campaign to create the appearance of a chemical disaster in Louisiana and an Ebola outbreak in Atlanta.130 There was real but limited harm in both of these cases, though the stories did not spread far because they lacked evidence and the facts were easy to check.

We will not always be so lucky as malicious attempts to spread panic grow. Deep fakes will prove especially useful for such disinformation campaigns, enhancing their credibility. Imagine if the Atlanta Ebola story had been backed by compelling fake audio appearing to capture a phone conversation with the head of the Centers for Disease Control and Prevention describing terrifying facts and calling for a cover-up to keep the public calm.

f. Undermining Diplomacy

Deep fakes will also disrupt diplomatic relations and roil international affairs, especially where the fake is circulated publicly and galvanizes public opinion. The recent Saudi-Qatari crisis might have been fueled by a hack that injected fake stories with fake quotes by Qatar's emir into a Qatari news site.131 The manipulator behind the lie could then further support the fraud with convincing video and audio clips purportedly gathered by and leaked from some unnamed intelligence agency.

A deep fake put into the hands of a state's intelligence apparatus may or may not prompt a rash action. After all, the intelligence agencies of the most capable governments are in a good position to make smart decisions about what weight to give potential fakes. But not every state has such capable institutions, and, in any event, the real utility of a deep fake for purposes of sparking an international incident lies in inciting the public in one or more states to believe that something shocking really did occur or was said. Deep fakes thus might best be used to box in a government through inflammation of relevant public opinion, constraining the government's options, and perhaps forcing its hand in some particular way. Recalling the concept of decisional chokepoints, for example, a well-timed deep fake calculated to inflame public opinion might be circulated during a summit meeting, making it politically untenable for one side to press its

---

129.   Cecilia Kang, *Hawaii Missile Alert Wasn't Accidental, Officials Say, Blaming Worker*, N.Y. TIMES (Jan. 30, 2018), https://www.nytimes.com/2018/01/30/technology/fcc-hawaii-missile-alert.html [https://perma.cc/4M39-C492].

130.   Adrian   Chen,   *The   Agency*,   N.Y.   TIMES   MAG.   (June   2,   2015), https://www.nytimes.com/2015/06/07/magazine/the-agency.html [https://perma.cc/DML3-6MWT].

131.   Krishnadev Calamur, *Did Russian Hackers Target Qatar?*, ATLANTIC (June 6, 2017), https://www.theatlantic.com/news/archive/2017/06/qatar-russian-hacker-fake-news/529359 [https://perma.cc/4QAW-TLY8] (discussing how Russian hackers may have planted a fake news story on a Qatari news site that falsely suggested that the Qatari Emir had praised Iran and expressed interest in peace with Israel).

agenda as it otherwise would have, or making it too costly to reach and announce some particular agreement.

g. *Jeopardizing National Security*

The use of deep fakes to endanger public safety or disrupt international relations can also be viewed as harming national security. But what else belongs under that heading?

Military activity—especially combat operations—belongs under this heading as well, and there is considerable utility for deep fakes in that setting. Most obviously, deep fakes have utility as a form of disinformation supporting strategic, operational, or even tactical deception. This is a familiar aspect of warfare, famously illustrated by the efforts of the Allies in Operation Bodyguard to mislead the Axis regarding the location of what became the D-Day invasion of June 1944.[132] In that sense, deep fakes will be (or already are) merely another instrument in the toolkit for wartime deception, one that combatants will both use and have used against them.

Critically, deep fakes may prove to have special impact when it comes to the battle for hearts and minds where a military force is occupying or at least operating amidst a civilian population, as was the case for the U.S. military for many years in Iraq and even now in Afghanistan. In that context, we have long seen contending claims about civilian casualties—including, at times, the use of falsified evidence to that effect. Deep fakes are certain to be used to make such claims more credible. At times, this will merely have a general impact in the larger battle of narratives. Nevertheless, such general impacts can matter a great deal in the long term and can spur enemy recruitment or enhance civilian support to the enemy. And, at times, it will spark specific violent reactions. One can imagine circulation of a deep-fake video purporting to depict American soldiers killing local civilians and seeming to say disparaging things about Islam in the process, precipitating an attack by civilians or even a host-state soldier or police officer against nearby U.S. persons.

Deep fakes pose similar problems for the activities of intelligence agencies. The experience of the United States since the Snowden leaks in 2013 demonstrates that the public, both in the United States and abroad, can become very alarmed about reports that the U.S. Intelligence Community has a particular capability, and this can translate into significant pressure to limit or abolish that capability both from an internal U.S. perspective and in terms of diplomatic relations. Whether those pressures resulted in changes that went too far in the case of the Snowden revelations is not our concern here. Our point is that this dynamic could be exploited if one wished to create distractions for an

---

132. Jamie Rubin, *Deception: the Other 'D' in D-Day*, NBC NEWS: THE ABRAMS REPORT (June 5, 2004), http://www.nbcnews.com/id/5139053/ns/msnbc-the_abrams_report/t/deception-other-d-d-day/#.WvQt5NMvyT8 [https://perma.cc/35HX-N7LN].

intelligence agency or generate conditions that would lead a society to limit what that agency is authorized to do. None of that would be easily done, but deep fakes make the prospect of a strategic operation to bedevil a competing state's intelligence services more plausible.[133]

The list of potential national security harms associated with deep fakes can go on, depending on one's definition of national security. In a recent report, the Belfer Center highlighted the national security implications of sophisticated forgeries.[134] An adversary could acquire real and sensitive documents through cyber-espionage and release the real documents along with forgeries. Deep-fake video and audio could be "leaked" to verify the forgeries. Foreign policy could be changed in response to convincing deep fakes and forgeries.[135]

h.   *Undermining Journalism*

As the capacity to produce deep fakes spreads, journalists increasingly will encounter a dilemma: when someone provides video or audio evidence of a newsworthy event, can its authenticity be trusted? That is not a novel question, but it will be harder to answer as deep fakes proliferate. News organizations may be chilled from rapidly reporting real, disturbing events for fear that the evidence of them will turn out to be fake.[136]

It is not just a matter of honest mistakes becoming more frequent: one can expect instances in which someone tries to trap a news organization using deep fakes. We already have seen many examples of "stings" pursued without the benefit of deep-fake technology.[137] Convincing deep fakes will make such stings more likely to succeed. Media entities may grow less willing to take risks in that

---

133.   In this context, it is interesting to note the success of the Shadow Brokers operation, which appears to have been a Russian effort not just to steal capabilities from NSA but to embarrass the NSA through a series of taunting public releases of those capabilities. There was also some degree of accompanying spin suggesting an interest in sowing doubt both in the U.S. and abroad about the wisdom of allowing the NSA to develop, keep, and use such capabilities in the first place. *See* Scott Shane, et al., *Security Breach and Spilled Secrets Have Shaken the N.S.A. to Its Core,* N.Y. TIMES (Nov. 12, 2017), https://www.nytimes.com/2017/11/12/us/nsa-shadow-brokers.html [https://perma.cc/WF6U-D4SV].

134.   GREG ALLEN & TANIEL CHAN, HARV. KENNEDY SCH. BELFER CTR. FOR SCI. AND INT'L AFF.,   ARTIFICIAL   INTELLIGENCE   AND   NATIONAL   SECURITY   (July   2017), https://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf [https://perma.cc/P4H5-QLVC].

135.   *Id.* at 34.

136.   Daniel Funke, *U.S. Newsrooms are 'Largely Unprepared' to Address Misinformation Online,* POYNTER (Nov. 14, 2017), https://www.poynter.org/news/us-newsrooms-are-largely-unprepared-address-misinformation-online [https://perma.cc/XUF4-8LLM].

137.   *See, e.g.*, Shawn Boburg, et al., *A Woman Approached the Post With Dramatic—and False—Tale About Roy Moore. She Appears to Be Part of Undercover Sting Operation*, WASH. POST (Nov. 27, 2017), https://www.washingtonpost.com/investigations/a-woman-approached-the-post-with-dramatic--and-false--tale-about-roy-moore-sje-appears-to-be-part-of-undercover-sting-operation/2017/11/27/0c2e335a-cfb6-11e7-9d3a-bcbe2af58c3a_story.html?utm_term=.6a4e98a07c2c [https://perma.cc/3TKD-27BP] (discussing an attempt to trick the Washington Post into running a false story about a woman claiming to have had sex as a teenager with and become pregnant by then-U.S. Senate candidate Roy Moore).

environment, or at least less willing to do so in timely fashion. Without a quick and reliable way to authenticate video and audio, the press may find it difficult to fulfill its ethical and moral obligation to spread truth.

i.   The Liar's Dividend: Beware the Cry of Deep-Fake News

We conclude our survey of the harms associated with deep fakes by flagging another possibility, one different in kind from those noted above. In each of the preceding examples, the harm stems directly from the use of a deep fake to convince people that fictional things really occurred. But not all lies involve affirmative claims that something occurred (that never did): some of the most dangerous lies take the form of denials.

Deep fakes will make it easier for liars to deny the truth in distinct ways. A person accused of having said or done something might create doubt about the accusation by using altered video or audio evidence that appears to contradict the claim. This would be a high-risk strategy, though less so in situations where the media is not involved and where no one else seems likely to have the technical capacity to expose the fraud. In situations of resource-inequality, we may see deep fakes used to escape accountability for the truth.

Deep fakes will prove useful in escaping the truth in another equally pernicious way. Ironically, liars aiming to dodge responsibility for their real words and actions will become more credible as the public becomes more educated about the threats posed by deep fakes. Imagine a situation in which an accusation is supported by genuine video or audio evidence. As the public becomes more aware of the idea that video and audio can be convincingly faked, some will try to escape accountability for their actions by denouncing authentic video and audio as deep fakes. Put simply: a skeptical public will be primed to doubt the authenticity of real audio and video evidence. This skepticism can be invoked just as well against authentic as against adulterated content.

Hence what we call the liar's dividend: this dividend flows, perversely, in proportion to success in educating the public about the dangers of deep fakes. The liar's dividend would run with the grain of larger trends involving truth skepticism. Most notably, recent years have seen mounting distrust of traditional sources of news. That distrust has been stoked relentlessly by President Trump and like-minded sources in television and radio; the mantra "fake news" has become an instantly recognized shorthand for a host of propositions about the supposed corruption and bias of a wide array of journalists, and a useful substitute for argument when confronted with damaging factual assertions. Whether one labels this collection of attitudes postmodernist or nihilist,[138] the

---

138.   For a useful summary of that debate, see Thomas B. Edsall, *Is President Trump a Stealth Postmodernist   or   Just   a   Liar?*,   N.Y.   TIMES   (Jan.   25,   2018), https://www.nytimes.com/2018/01/25/opinion/trump-postmodernism-lies.html [https://perma.cc/DN7F-AEPA].

fact remains that it has made substantial inroads into public opinion in recent years.

Against that backdrop, it is not difficult to see how "fake news" will extend to "deep-fake news" in the future. As deep fakes become widespread, the public may have difficulty believing what their eyes or ears are telling them—even when the information is real. In turn, the spread of deep fakes threatens to erode the trust necessary for democracy to function effectively.139

The combination of *truth* decay and *trust* decay accordingly creates greater space for authoritarianism. Authoritarian regimes and leaders with authoritarian tendencies benefit when objective truths lose their power.140 If the public loses faith in what they hear and see and truth becomes a matter of opinion, then power flows to those whose opinions are most prominent—empowering authorities along the way.141

Cognitive bias will reinforce these unhealthy dynamics. As Part II explored, people tend to believe facts that accord with our preexisting beliefs.142 As research shows, people often ignore information that contradicts their beliefs and interpret ambiguous evidence as consistent with their beliefs.143 People are also inclined to accept information that pleases them when given the choice.144 Growing appreciation that deep fakes exist may provide a convenient excuse for motivated reasoners to embrace these dynamics, even when confronted with information that is in fact true.

III.

WHAT CAN BE DONE? EVALUATING TECHNICAL, LEGAL, AND MARKET RESPONSES

What can be done to ameliorate these harms? Part III reviews various possibilities. To start, we explore the prospects for technological solutions that would facilitate the detection and debunking of deep fakes. We then describe

---

139.   The Edelman Trust Barometer, which measures trust in institutions around the world, recorded a drop of nine points in the Trust Index for the United States from 2017 to 2018. Even among the informed public, the US dropped from a Trust Index of 68 to 45. 2018 EDELMAN TRUST BAROMETER GLOBAL REPORT 7 (2018), https://www.edelman.com/sites/g/files/aatuss191/files/2018-10/2018_Edelman_Trust_Barometer_Global_Report_FEB.pdf [https://perma.cc/Z26M-GQ2A].

140.   MILES BRUNDAGE ET AL., THE MALICIOUS USE OF ARTIFICIAL INTELLIGENCE: FORECASTING, PREVENTION, AND MITIGATION 46 (2018) https://www.eff.org/files/2018/02/20/malicious_ai_report_final.pdf [https://perma.cc/K2KT-XVZQ].

141.   *Id.*

142.   *See generally* Michela Del Vicario et al., *Modeling Confirmation Bias and Polarization*, 7 SCIENTIFIC REPORTS no. 40, 391 (2017) (assessing models that describe polarization effects relating to cognitive biases).

143.   *See generally* Constanza Villarroel et al., *Arguing Against Confirmation Bias: The Effect of Argumentative Discourse Goals on the Use of Disconfirming Evidence in Written Argument*, 79 INT'L J. EDUC. RES. 167 (2016) (demonstrating impact of biases on belief formation).

144.   *See generally* Shanto Iyengar et al., *Selective Exposure to Campaign Communication: The Role of Anticipated Agreement and Issue Public Membership*, 70 J. POL. 186 (2008) (examining impact of bias models in the context of political campaign information).

current and potential proposals for criminal and civil liability. With law in mind, we discuss the role of regulators and identify ways in which the government might respond to deep fakes. In the shadow of these possibilities, we anticipate new services the market might spawn to protect individuals from harm associated with deep fakes—and the considerable threat to privacy such services themselves might entail.

A.   *Technological Solutions*

Technology has given us deep fakes – but might it also provide us with a capacity for debunking them and limiting their harmful potential? An efficient and generally effective method for rapid detection of deep fakes would go far toward resolving this topic as a matter of pressing public-policy concern. Unfortunately, the challenges are daunting. For example, detection software would have to keep pace with innovations in deep-fake technology to retain efficacy. Moreover, if such technology existed and could be deployed through social media platforms, it would only reduce the systemic harms described above, but by no means eliminate them. Such developments might not protect individuals from deep fakes involving narrow or even isolated distribution.[145] Further, detection software might not disabuse certain people's faith in deep fakes, especially those under the profound sway of cognitive bias. At the least though, the impact of harmful deep fakes might be cabined while beneficial uses could continue unabated.

At any rate, it is far from clear that such technology will emerge in the near future. There are a number of projects—academic and corporate—aimed at creating counterfeit-proof systems for authenticating content or otherwise making it easier to confirm credible provenance.[146] Such systems, however, are tailored to particular products rather than video or audio technologies generally. They will therefore have only limited use until one program becomes ubiquitous and effective enough for dominant platforms to incorporate them into their content-screening systems—and, indeed, to make use of them mandatory for posting. Additionally, these systems will have to withstand users' efforts to bypass them.

For now, we are left to seek a generally applicable technology that can detect manipulated content without an expectation that the content comes with

---

145.   GIF hosting company Gyfcat has developed and trained AI to spot fraudulent videos. Project Maru, as they call it, can spot deep-fake videos because in many frames, the faces aren't perfectly rendered. They have also developed Project Angora, which "mask[s]" the face of a possible deep fake and searches the Internet to see if the body and background footage exist elsewhere. *See* Louise Matsakis, *Artificial Intelligence is Now Fighting Fake Porn*, WIRED (Feb. 14, 2018) https://www.wired.com/story/gfycat-artificial-intelligence-deepfakes [https://perma.cc/PX4N-VZJY].

146.   For examples of provenance technologies in development, see Dia Kayyali, *Set Your Phone to ProofMode*, WITNESS, https://blog.witness.org/2017/04/proofmode-helping-prove-human-rights-abuses-world [https://perma.cc/GB6M-KQPF] (describing the concept of a metadata-rich "ProofMode" app for Android devices).

an internal certification. Professor Hany Farid, the pioneer of PhotoDNA, a technology that identifies and blocks child pornography, warns: "We're decades away from having forensic technology that . . . [could] conclusively tell a real from a fake . . . If you really want to fool the system you will start building into the deepfake ways to break the forensic system."[147] The defense, in short, is currently faring poorly in the deep-fake technology arms race.

As problems associated with deep fakes begin to accumulate, we might expect developments that could alter the current balance of power between technologies that create deep fakes and those that detect them. For example, growing awareness of the problem might produce the conditions needed for grantmaking agencies like the National Science Foundation and the Defense Advanced Research Projects Agency (DARPA) to begin steering funds toward scalable detection systems that can be commercialized or even provided freely. DARPA has an initial project in the form of a contest pitting GAN methods for generating deep fakes against would-be detection algorithms. The DARPA project manager is skeptical about the prospects for detection, however, given current technical capacities.[148]

Emerging market forces might encourage companies to invest in such capabilities on their own or in collaboration with each other and with academics (a possibility that we revisit below). For now, however, it would be foolish to trust that technology will deliver a debunking solution that is scalable and reliable enough to minimize the harms deep fakes might cause.

B.   Legal Solutions

If technology alone will not save us, might the law? Would a combination of criminal and civil liability meaningfully deter and redress the harms that deep fakes seem poised to cause? We examine the possibilities under existing and potential law.

1.   Problems with an Outright Ban

No current criminal law or civil liability regime bans the creation or distribution of deep fakes. A threshold question is whether such a law would be normatively appealing and, if so, constitutionally permissible.

A flat ban is not desirable because digital manipulation is not inherently problematic. Deep fakes exact significant harm in certain contexts but not in all. A prohibition of deep fakes would bar routine modifications that improve the

---

147.   *See* Matsakis, *supra* note 145 (quoting Prof. Hany Farid in reference to "fake porn").

148.   *See* Will Knight, *The U.S. Military Is Funding an Effort to Catch Deepfakes and Other AI Trickery*, MIT TECH. REV. (May 23, 2018), https://www.technologyreview.com/s/611146/the-us-military-is-funding-an-effort-to-catch-deepfakes-and-other-ai-trickery [https://perma.cc/7RD7-5CMJ] ("'Theoretically, if you gave a GAN all the techniques we know to detect it, it could pass all of those techniques,' says David Gunning, the DARPA program manager in charge of the project. 'We don't know if there's a limit. It's unclear.'").

clarity of digital content. It would chill experimentation in diverse fields, from history and science to art and education.

Crafting a law prohibiting destructive applications of deep-fake technology while excluding beneficial ones would be difficult, but perhaps not impossible. For example, what if a law required proof of a deep-fake creator's intent to deceive and evidence of serious harm as a way to reduce concerns about chilling public discourse? Under such a proposal, concerns over speech still remain. The very existence of a general prohibition of deep fakes, even with those guardrails, would cast a significant shadow, potentially diminishing expression crucial to self-governance and democratic culture. The American free speech tradition warns against government having the power to pick winners and losers in the realm of ideas because it will "tend to act on behalf of the ideological powers that be."[149] As James Weinstein notes, we should be especially wary of entrusting government officials with the power to determine the veracity of factual claims "made in the often highly ideological context of public discourse . . . ."[150] A deep-fakes ban would raise the specter of penalties for parodies of would-be or current office holders.

Although self-serving prosecutions are not inevitable, they are a real possibility.[151] Dislike of minority or unpopular viewpoints, combined with ambiguity surrounding a deep-fake creator's intent, might result in politicized enforcement.[152] This might inhibit engagement in political discourse

---

149. Frank I. Michelman, *Conceptions of Democracy in American Constitutional Argument: The Case of Pornography Regulation*, 56 TENN. L. REV. 291, 302 (1989); *see also* Thomas v. Collins, 323 U.S. 516, 545 (1945) (Jackson, J., concurring) ("[T]he forefathers did not trust any government to separate the true from the false for us."). Justice Oliver Wendell Holmes cautioned against the human inclination to silence opinions that we dislike. *See* Abrams v. United States, 250 U.S. 616, 630 (1919) (Holmes, J., dissenting) ("[W]e should be eternally vigilant against attempts to check the expression of opinions that we loathe . . . ."). "Persecution for the expression of opinions[,]" he wrote, is "perfectly logical . . . [i]f you have no doubt of your premises or your power and want a certain result with all your heart . . . ." *Id.* Holmes offered against this certainty, and power's tendency to sweep away disagreement, a principle of epistemic doubt that is a defining hallmark of First Amendment law. *See id.*

150. James Weinstein, *Climate Change Disinformation, Citizens Competence, and the First Amendment*, 89 U. COLO. L. REV. 341, 351 (2018).

151. Indeed, public officials recently have called for a rethinking of libel laws. Alex Pappas, *Trump: 'Our Current Libel Laws Are a Sham'*, FOX NEWS (Jan. 10, 2018), https://www.foxnews.com/politics/trump-our-current-libel-laws-are-a-sham [https://perma.cc/AHM4-UN6G]; Gregg Re, *Clarence Thomas Backs Trump's Call for Changing Defamation Law for Easing Suits Against the Media*, FOX NEWS (Feb. 19, 2019), https://www.foxnews.com/politics/clarence-thomas-calls-for-easing-defamation-suits-by-politicians-like-trump [https://perma.cc/RN42-DFCK]. Although this suggestion may seem untenable given the U.S. commitment to robust and wide-open debate on public issues, it animates concerns about partisan enforcement.

152. Weinstein, *supra* note 150, at 351 ("There is even greater reason to distrust the ability of government officials to fairly and accurately determine the speaker's state of mind in making allegedly false statement."). James Weinstein explains that "government officials hostile to the speaker's point of view are more likely to believe that the speaker knew that the statement was false, while officials who share the speaker's ideological perspective will be more likely to find that any misstatement of fact was an innocent one." *Id.*

specifically, and in democratic culture more generally.153 The "'risk of censorious selectivity by prosecutors'" [will] . . . distort perspectives made available" to the public.154 It is far better to forego an outright ban of deep fakes than to run the risk of its abuse.

Even if these normative concerns could be overcome, it is unlikely that a flat ban on deep fakes could withstand constitutional challenge. Deep fakes implicate freedom of expression, even though they involve intentionally false statements.155 In the landmark 1964 decision *New York Times v. Sullivan*,156 the Supreme Court held that false speech enjoys constitutional protection insofar as its prohibition would chill truthful speech.157

In 2012, in *United States v. Alvarez*,158 the Court went even further. In the plurality and concurring opinions, the Court concluded that "falsity alone" does not remove expression from First Amendment protection.159 As Justice Kennedy's plurality noted, falsehoods generally warrant protection because they inspire rebuttal and "reawaken respect" for valuable ideas in public discourse.160 Central to this point is faith in the public's willingness to counter lies and engage in reasoned discourse.

While all nine Justices agreed that the harmful effect of false factual statements could be regulated, they differed in the particulars.161 The plurality opinion took the position that false statements can be proscribed if the speakers intended to cause "legally cognizable harm" of a kind traditionally understood as falling outside the First Amendment's protection.162 The concurrence posited that a law aimed at regulating harm-causing falsehoods may be permissible if it

---

153. For Jack Balkin's influential theory of free speech grounded in participation in democratic culture, see, for example, Jack M. Balkin, *Cultural Democracy and the First Amendment*, 110 NW. U. L. REV. 1053, 1072 (2016) (arguing that key to free society is ability to engage in meaning making and creation of culture).

154. Weinstein, *supra* note 150, at 360 (quoting United States v. Alvarez, 567 U.S. 709, 736 (2012) (Breyer, J., concurring)).

155. *See generally* Lewis Sargentich, Note, *The First Amendment Overbreadth Doctrine,* 83 HARV. L. REV. 844, 845 (1970) (describing a judicial presumption against statutes that curtail a broad array of expressive activity).

156. N.Y. Times v. Sullivan, 376 U.S. 254 (1964).

157. *Id.* at 264.

158. *Alvarez*, 567 U.S. 709 (plurality opinion).

159. *Id.* at 719. For a superb discussion of the constitutional significance of lies in the aftermath of *Alvarez*, see Alan K. Chen & Justin Marceau, *High Value Lies, Ugly Truths, and the First Amendment*, 68 VAND. L. REV. 1435, 1440–54 (2015). *See generally* Geoffrey R. Stone, *Kenneth Karst's Equality as the Central Principle in the First Amendment*, 75 U. CHI. L. REV. 37, 43 (2008) (discussing a "two-level" theory of the First Amendment: one that treats high value speech with stringent protections, and a second tier of speech that falls outside the First Amendment's coverage).

160. *Alvarez*, 567 U.S. at 719, 722 ("Indeed, the outrage and contempt expressed for respondent's lies can serve to reawaken and reinforce the public's respect for the Medal, its recipients, and its high purpose.").

161. *Id.* at 719 (plurality opinion); *id.* at 731–34 (Breyer, J., concurring); *id.* at 750 (Alito, J. dissenting).

162. *Id.* at 719, 725 (plurality opinion).

does not disproportionately damage First Amendment interests.163 The dissent would have denied First Amendment protection to false factual statements that inflict harm and serve no legitimate purpose.164 The court reached consensus that regulation of false statements involving history, politics, literature, and other matters of public concern requires strict scrutiny review.165

The opinions in Alvarez, taken together, would seem to preclude a sweeping ban on deep fakes while leaving considerable room for carefully tailored prohibitions of certain harmful deep fakes. As the plurality underscored in *Alvarez*, certain categories of speech are not covered by the First Amendment due to their propensity to bring about serious harms and their slight contribution to free speech values.166 Some deep fakes will fall into those categories and thus could be subject to regulation. This includes defamation of private persons, fraud, true threats, and the imminent-and-likely incitement of violence.167 Speech integral to criminal conduct like extortion, blackmail, and perjury has long been understood to enjoy no First Amendment protection.168

Consider as an illustration laws banning the impersonation of government officials (such as law enforcement officers or agency officials). As Helen Norton insightfully explains, these statutes are "largely uncontroversial as a First Amendment matter in great part because they address real (if often intangible) harm to the public as well as to the individual target."169 Lies about the source of speech—whether a public official is actually speaking—do not serve free speech values.170 Quite the opposite, they deny listeners the ability to assess the quality and credibility of the speech, undermining democratic self-governance and the search for truth.171 From a normative perspective, therefore, a surgical approach

---

163.   *Id.* at 737 (Breyer, J. concurring).

164.   *Id.* at 750 (Alito, J. dissenting).

165.   *Id.* at 722 (Kennedy, J. plurality); *id.* at 734 (Breyer, J. concurring); *id.* at 751 (Alito, J., dissenting). For an insightful exploration of *Alvarez* and its implications for the regulation of deep fakes, see Marc Jonathan Blitz, *Lies, Line Drawing, and (Deep) Fake News*, 71 OKLA. L. REV. 59, 110 (2018) (arguing that government should have greater power to regulate forgeries than the malicious statement of false facts); Marc Jonathan Blitz, Deep Fakes and Other Non-Testimonial Falsehoods: When Is Belief Manipulation (Not) First Amendment Speech? (Apr. 18, 2019) (unpublished manuscript) (on file with authors) (arguing that deep fakes may fall outside of First Amendment coverage because they arguably amount to non-testimonial evidence and change perceptions of the world around them, especially where government seeks to require disclosure that something is a deep fake).

166.   *See generally* CITRON, HATE CRIMES IN CYBERSPACE, *supra* note 40, at 199–218 (discussing narrow categories of low-value speech accorded less rigorous protection or no protection under First Amendment analysis).

167.   *See* Chen & Marceau, *supra* note 159, at 1480–91.

168.   CITRON, HATE CRIMES IN CYBERSPACE, *supra* note 40, at 203–05 (explaining that crime-facilitating speech does not enjoy First Amendment protection in context of cyber stalking).

169.   Helen Norton, *Lies to Manipulate, Misappropriate, and Acquire Government Power*, *in* LAW AND LIES 143, 170 (Austin Sarat ed., 2015) [hereinafter Norton, *Lies to Manipulate*].

170.   *Id.* at 168. We are grateful to both Helen Norton and Marc Blitz who generously spent time talking to us about the doctrinal and theoretical free speech issues raised in regulating deep fakes.

171.   Helen Norton, *(At Least) Thirteen Ways of Looking at Election Lies*, 71 OKLA. L. REV. 117, 131 (2018).

to criminal and civil liability may result in a more attractive balance of costs and benefits than a deep-fake ban perspective. And so we turn now to a discussion of specific possibilities, starting with civil liability.

2. *Specific Categories of Civil Liability*

Given that deep fakes cannot and should not be banned on a generalized basis, the question remains whether their creators and distributors in particular contexts should be subject to civil liability for the harms they cause. This section reviews relevant existing laws and possible improvements.

a. *Threshold Obstacles*

Before reviewing the prospects for particular theories of liability, we note two threshold problems.

The first involves attribution. Civil liability cannot ameliorate harms caused by deep fakes if plaintiffs cannot tie them to their creators. The attribution problem arises in the first instance because the metadata relevant for ascertaining a deep fake's provenance might be insufficient to identify the person who generated it. It arises again when the creator or someone else posts a deep fake on social media or otherwise injects it into the marketplace of information. A careful distributor of a deep fake may take pains to be anonymous, including but not limited to using technologies like Tor.[172] When these technologies are employed, the IP addresses connected to posts may be impossible to find and trace back to the responsible parties.[173] In such cases, a person or entity aggrieved by a deep fake may have no practical recourse against its creator, leaving only the possibility of seeking a remedy from the owner of platforms that enabled circulation of the content.

A second obstacle arises when the creator of the deep fake—or the platform circulating it—is outside the United States and thus beyond the effective reach of US legal process, or in a jurisdiction where local legal action is unlikely to be effective. Therefore, even if attribution is known, it still may be impossible to use civil remedies effectively. While limitations of civil liability exist in many settings, the global nature of online platforms makes it a particular problem in the deep-fake context.

Moreover, regardless of whether perpetrators can be identified or reside in the US, civil suits are expensive. Victims usually bear the heavy costs of bringing civil claims and may be hesitant to initiate lawsuits if deep-fake generators are

---

172. *See* CITRON, HATE CRIMES IN CYBERSPACE, *supra* note 40, at 142–43 (arguing that law has difficulty communicating norms, deterring unlawful activity, or redressing injuries if defendants have used anonymizing technologies that make it difficult to identity them).

173. Citron, *Cyber Civil Rights*, *supra* note 46, at 117 (explaining that claims cannot be pressed against cyber stalkers if websites hosting their abuse fails to track IP addresses).

effectively judgment-proof.174 Worse, the "Streisand Effect" is likely to overhang the decision to sue when the deep fake is embarrassing or reputationally harmful. Lawsuits attract publicity; unless the victim is permitted to sue under a pseudonym, filing a claim may exacerbate the victim's harm.175

b. *Suing the Creators of Deep Fakes*

Threshold attribution and liability hurdles are not always fatal for would-be plaintiffs. When a victim decides to sue the creator of a deep fake, several bodies of law come into play, including intellectual property and tort law.

First, consider copyright law. Some deep fakes exploit copyrighted content, opening the door to monetary damages and a notice-and-takedown procedure that can result in removal of the offending content.176 A copyright owner is the person who took a photograph. Thus, if a deep fake involves a photo that the victim took of herself, the victim might have a copyright claim against the creator of the deep fake.177

The prospects for success, however, are uncertain. A court will have to determine whether the deep fake is a "fair use" of the copyrighted material, intended for educational, artistic, or other expressive purposes. Whether the fake is sufficiently transformed from the original to earn fair use protection is a highly fact-specific inquiry for which a judicial track record does not yet exist.178

Tort law also includes concepts that could be used to address deep-fake scenarios. Most obviously, victims can sue for defamation. Where the alleged defamation concerns private individuals rather than public figures, states may permit plaintiffs to prevail based on a showing that the falsehood was made negligently.179 Public officials and public figures are subject to a higher requirement of showing clear and convincing evidence of actual malice— knowledge or reckless disregard for the possibility that the deep fakes were

---

174.   *See* CITRON, HATE CRIMES IN CYBERSPACE, *supra* note 40, at 122 (exploring limits of civil law in redressing injuries resulting from cyber stalking).

175.   Mike Masnick coined the phrase "the Streisand Effect" in *Techdirt* in 2005. Mike Masnick, *Since When Is It Illegal to Just Mention a Trademark Online?*, TECHDIRT (Jan. 5, 2005), https://www.techdirt.com/articles/20050105/0132239_F.shtml [https://perma.cc/XR42-G9BX].

176.   *See* Derek E. Bambauer, *Exposed*, 98 MINN. L. REV. 2025, 2065–67 (2014) (discussing the removal of copyright-infringing material).

177.   CITRON, HATE CRIMES IN CYBERSPACE, *supra* note 40, at 122 (explaining that someone can sue for copyright of their own image only if they took the photos themselves); *see also* Megan Farokhmanesh, *Is It Legal to Swap Someone's Face Into Porn Without Consent?*, VERGE (Jan. 30, 2018), https://www.theverge.com/2018/1/30/16945494/deepfakes-porn-face-swap-legal [https://perma.cc/TH4N-YUJV] (quoting Eric Goldman).

178.   *Compare* David Greene, *We Don't Need New Laws for Faked Videos, We Already Have Them*, EFF BLOG (Feb. 13, 2018), https://www.eff.org/deeplinks/2018/02/we-dont-need-new-laws-faked-videos-we-already-have-them [https://perma.cc/KEG4-73L3] (noting that copyright claims may address deep fakes subject to fair use objections) *with* Jesse Lempel, *Combatting Deep Fakes Through the Right of Publicity*, LAWFARE (Mar. 30, 2018), https://www.lawfareblog.com/combatting-deep-fakes-through-right-of-publicity [https://perma.cc/6TPH-98S9].

179.   *See Gertz v. Welch*, 418 U.S. 323, 343–46 (1974).

false.180 In addition to defamation, the closely related tort of placing a person in a "false light"—or recklessly creating a harmful and false implication about someone in a public setting—has clear potential for the deep fake context.181

Victims may also sue in tort for intentional infliction of emotional distress. This requires proof of "extreme and outrageous conduct."182 Creating and circulating humiliating content like deep-fake sex videos would likely amount to "extreme and outrageous conduct" because it falls outside the norms of decency by most accounts.183

Another prospect is the "right of publicity" in tort law, which permits compensation for the misappropriation of someone's likeness for commercial gain.184 The commercial-gain element sharply limits the utility of this model: the harms associated with deep fakes do not typically generate direct financial gain for their creators.185 This is likely true, for example, of deep fakes posted to harm rivals or ex-lovers. Only in core cases, such as a business using deep-fake technology to make it seem a particular person endorsed their product or service, might this approach prove useful in stemming abuse. Further, the expressive value of some deep fakes may constitute a further hurdle to liability; courts often dismiss right of publicity claims concerning newsworthy matters on free-speech grounds.186

Other privacy-focused torts seem relevant at first blush, yet are a poor fit on close inspection.187 The "public disclosure of private fact" tort, for example, allows individuals to recover for publication of private, "non-newsworthy" information that would highly offend the reasonable person.188 While deep fakes may meet the offense standard, using a person's face in a deep-fake video does not amount to the disclosure of *private* information if the source image was publicly available.189 The "intrusion-on-seclusion" tort is likewise ill-suited to

---

180. *See* RESTATEMENT (SECOND) OF TORTS § 559 (AM. LAW INST. 1969); *see also* CITRON, HATE CRIMES IN CYBERSPACE, *supra* note 40, at 121, 132–34 (explaining the reach of defamation law in cases involving private individuals and public figures).

181. CITRON, HATE CRIMES IN CYBERSPACE, *supra* note 40, at 121, 132–34.

182. *See* CITRON, HATE CRIMES IN CYBERSPACE, *supra* note 40, at 133–34, 140–41 (explaining that emotional distress claims are warranted for online abuse that is targeted, cruel, and reliant on sensitive embarrassing information, including nude photos).

183. *See* Benjamin C. Zipursky, Snyder v. Phelps, *Outrageousness, and the Open Texture of Tort Law,* 60 DEPAUL L. REV. 473 (2011).

184. *See generally* JENNIFER F. ROTHMAN, RIGHT OF PUBLICITY: PRIVACY REIMAGINED FOR A PUBLIC WORLD (2018) (summarizing the history of the right of publicity tort).

185. *See generally* Lempel, *supra* note 178 (discussing how right to publicity claims would likely only succeed against misappropriations intended for commercial gain).

186. *See generally* ROTHMAN, *supra* note 184.

187. *See generally* Danielle Keats Citron, *Mainstreaming Privacy Torts*, 98 CALIF. L. REV. 1805, 1811–14 (2010) [hereinafter Citron, *Mainstreaming Privacy Torts*] (exploring the limited application of privacy torts to twenty-first century privacy harms).

188. DANIEL J. SOLOVE & PAUL M. SCHWARTZ, PRIVACY LAW FUNDAMENTALS 47 (4th ed. 2017).

189. *See id.* at 42, 49. One of us (Citron) has explored the limits of privacy torts in context of deep-fake sex videos. Citron, *Sexual Privacy*, *supra* note 93, at 1933–35.

the deep-fake scenario. It narrowly applies to defendants who "intruded into a private place, or . . . invaded a private seclusion that the plaintiff has thrown about his person or affairs."[190] Deep-fakes usually will not involve invasions of spaces (either physical or conceptual like email inboxes) in which individuals have a reasonable expectation of privacy.

Therefore, current options for imposing liability on creators of deep fakes have mixed potential. Civil liability is most robust in relation to defamation, false light, and intentional infliction of emotional distress, with more limited prospects for copyright infringement and right of publicity claims.

c.   Suing the Platforms

It will be challenging to achieve individualized accountability for harmful deep fakes, but creators are not the only parties that might bear responsibility. Given the key role that content platforms play in enabling the distribution of deep fakes, and the fact that creators of harmful deep fakes in some cases may be difficult to find and deter, the most efficient and effective way to mitigate harm may be to impose liability on platforms.[191] In some contexts, this may be the only realistic possibility for deterrence and redress.

Online platforms already have an incentive to screen content, thanks to the impact of moral suasion, market dynamics, and political pressures.[192] They do not currently face significant civil liability risk for user-generated content, however, for the reasons explained below.

In 1996, Congress provided platforms with a liability shield in the form of Section 230 of the Communications Decency Act (CDA). The law provided an immunity from liability to online platforms for hosting harmful content, albeit with an exception for content that violates federal criminal law, the Electronic Communications Privacy Act, and intellectual property law.[193]

Section 230 protects platforms in important ways. First, consider a situation in which an online platform displays content that links to another source (such as a news article or blog post) or is user-generated (such as a customer review

---

190.   RESTATEMENT (SECOND) OF TORTS § 652B, cmt. c (AM. LAW INST. 1969).

191.   *See* Citron, *Mainstreaming Privacy Torts*, *supra* note 187, at 1839–40.

192.   *See* Citron, *Extremist Speech*, *supra* note 46, at 1047–48; Citron, *Sexual Privacy*, *supra* note 93, at 1955–58 (examining Facebook's developing strategy to address nonconsensual pornography in response to victims' concerns brought to the company by advocacy groups such as the Cyber Civil Rights Initiative); Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1616–30 (2018); *see also* CITRON, HATE CRIMES IN CYBERSPACE, *supra* note 40, at 227–30 (exploring how and why content platforms moderate harmful content); Danielle Keats Citron & Helen Norton, *Intermediaries and Hate Speech: Fostering Digital Citizenship For Our Information Age*, 91 B.U. L. REV. 1435, 1454–59 (2011) (describing varied steps platforms have taken to moderate digital hate, motivated by moral, business, and other instrumental concerns). One of us (Citron) is the Vice President of the Cyber Civil Rights Initiative and has advised social media platforms about concerns of cyber stalking victims for the past ten years, importantly without compensation.

193.   47 U.S.C. § 230(c)(1) (2012). *See generally* Citron & Wittes, *supra* note 46.

posted on Yelp). Now, imagine that the content is defamatory or otherwise actionable. Can the plaintiff sue the online platform that helped it see the light of day? Not under Section 230. Section 230(c)(1) expressly forbids treating the platform as a "publisher" or "speaker" of someone else's problematic content. As courts have interpreted Section 230, online platforms enjoy immunity from liability for user-generated content even if they deliberately encouraged the posting of that content.[194]

Next, consider a situation in which an online platform decides not to allow users to post whatever they wish, but to instead screen and block certain harmful content. Might the act of filtering become the basis of liability? If so, platforms might be loath to do any screening at all. Section 230(c)(2) was meant to remove the disincentive to self-regulation that liability otherwise might produce.[195] Simply put, it forbids civil suits against platforms based on the good-faith act of filtering to screen out offensive content, whether in the nature of obscenity, harassment, violence, or otherwise.[196]

In crafting Section 230, the bill's sponsors thought they were devising a safe harbor for online service providers that would enable the growth of the then-emerging "Internet."[197] Representative Chris Cox, for example, became interested after reading about a trial court decision holding Prodigy, an online services company, liable as a publisher of defamatory comments because it tried but failed to filter profanity on its bulletin boards.[198] A key goal of the legislation

---

194. *See* Citron & Wittes, *supra* note 46, at 408–09 (laying out judicial decisions interpreting Section 230 that have produced sweeping immunity from liability for user-generated content, including for sites that encourage users to post illegal content and sites that knowingly and deliberately repost illegal content); *see also* CITRON & JURECIC, PLATFORM JUSTICE, *supra* note 46 (same). In one example, Michael Herrick sued Grindr, a dating app, after the site refused to remove a user who was impersonating him on the app, sharing his nude images, claiming he had rape fantasies, and providing his home address. Herrick v. Grindr, 306 F. Supp. 3d 579, 585–86 (S.D.N.Y. 2018). More than 1,000 men came to Herrick's home demanding sex. *Id.* at 588. Grindr refused to address Herrick's large number of complaints. *Id.* The district court dismissed the case on Section 230 grounds, which the Second Circuit affirmed in a summary order. *Id.*; Herrick v. Grindr, 765 Fed. Appx. 586 (2d Cir. 2019).

195. 47 U.S.C. § 230(c)(2); Citron & Wittes, *supra* note 46, at 406 (explaining that Section 230(c)(2) provides broad protections for good-faith over-screening of content).

196. *See* 47 U.S.C. § 230(c)(2) ("No provider or user of an interactive computer service shall be held liable on account of . . . any action voluntarily taken in good faith to restrict access to . . . material that the provide or user considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable . . . .").

197. Danielle Keats Citron, *Section 230's Challenge to Civil Rights and Civil Liberties*, KNIGHT FIRST AMEND. INST. [hereinafter Citron, *Section 230's Challenge*], https://knightcolumbia.org/content/section-230s-challenge-civil-rights-and-civil-liberties [https://perma.cc/MHN7-JXZJ] (describing history of § 230 and recent developments). *See generally* CITRON, HATE CRIMES IN CYBERSPACE, supra note 40, at 170. For an illuminating explanation of the cases that prompted the adoption of Section 230 and its broad interpretation, see generally JEFF KOSSEFF, THE TWENTY-SIX WORDS THAT CREATED THE INTERNET (2019).

198. The firm in question happens to have been the one that is the subject of the film *Wolf of Wall Street*. *See* Alina Selyukh, *Section 230: A Key Legal Shield for Facebook, Google is About to Change*, NPR MORNING EDITION (Mar. 21, 2018),

was to help "clean up" the Internet by making it easier for willing platforms to filter out offensive material, removing the risk that doing so would incur civil liability by casting them in a publisher's role.199

At the time, sponsors Senators James Exon and Slade Gorton sought to combat online pornography and make the "Internet" safe for kids.200 Representatives Cox and Ron Wyden, another sponsor, argued that, if "this amazing new thing—the Internet—[was] going to blossom," companies should not be "punished for *trying* to keep things clean."201

This intent is clear in the language of Section 230(c)(2), which expressly concerns platforms engaged in "good faith" editorial activity involving the blocking and filtering of offensive user-posted content. The speaker and publisher liability provision of Section 230, however, lacks this narrowing language and has become a foundation for courts to interpret Section 230 immunity broadly.202

No doubt, Section 230's immunity provision has been beneficial for digital expression and democratic culture. It has provided breathing room for the development of online services and innumerable opportunities for speech and discourse.203 Its supporters contend that without immunity, search engines, social networks, and microblogging services might not have emerged.204 We agree; the fear of publisher liability would likely have inhibited the Internet's early growth.205

However, an overbroad reading of Section 230 has "given online platforms a free pass to ignore illegal activities, to deliberately repost illegal material, and to solicit unlawful activities while ensuring that abusers cannot be identified."206 The permissive interpretation of Section 230 eliminates "incentives for better

---

https://www.npr.org/sections/alltechconsidered/2018/03/21/591622450/section-230-a-key-legal-shield-for-facebook-google-is-about-to-change [https://perma.cc/S9K9-GX47].

199.	*See* Citron & Wittes, *supra* note 46, at 405–06.

200.	*See* S. REP. NO. 104-23, at 59 (1995). Key provisions criminalized the transmission of indecent material to minors.

201.	Selyukh, *supra* note 198 (quoting Cox); *see* CITRON, HATE CRIMES IN CYBERSPACE, *supra* note 40, at 170–72 (describing the original purpose of Section 230's immunity provision).

202.	*See* Citron, *Cyber Civil Rights*, *supra* note 46, at 121–23; Citron & Wittes, *supra* note 46, at 408–10. In the landmark *Reno v. ACLU* decision, the Supreme Court struck down the CDA's blanket restrictions on Internet indecency under the First Amendment. *See* Reno v. ACLU, 521 U.S. 844 (1997). Online expression was too important to be limited to what government officials think is fit for children. *Id.* at 875. Section 230's immunity provision, however, was left intact.

203.	Citron & Wittes, *supra* note 46, at 413.

204.	CITRON, HATE CRIMES IN CYBERSPACE, *supra* note 40, at 171. For some of the most insightful work on the significance of Section 230's immunity provision, see the work of Daphne Keller, Jeff Kosseff, and Mike Godwin. *See, e.g.*, MIKE GODWIN, CYBER RIGHTS: DEFENDING FREE SPEECH IN THE DIGITAL AGE 319–54 (2003); KOSSEFF, *supra* note 197; Daphne Keller, *Toward a Clearer Conversation about Platform Liability,* KNIGHT FIRST AMEND. INST., https://knightcolumbia.org/content/toward-clearer-conversation-about-platform-liability [https://perma.cc/YSS5-WHJG].

205.	*Id.*

206.	Citron & Wittes, *supra* note 46, at 413.

behavior by those in the best position to minimize harm."[207] The results have been two-fold. On one hand, the law has created an open environment for hosting and distributing user-generated online content. On the other, it has generated an environment in which it is exceptionally hard to hold providers accountable, even in egregious circumstances involving systematic disinformation and falsehoods.[208]

Courts have extended the immunity provision to a remarkable array of scenarios. They include instances where a provider republished content knowing it violated the law;[209] solicited illegal content while ensuring that those responsible could not be identified;[210] altered its user interface to ensure that criminals could were not caught;[211] and sold dangerous products.[212] In this way, Section 230 has evolved into a super-immunity that, among other things, prevents the best-positioned entities to respond to most harmful content. This would have seemed absurd to the CDA's drafters.[213] The law's overbroad interpretation means that platforms have no liability-based reason to take down illicit material, and that victims have no legal leverage to insist otherwise.[214] Rebecca Tushnet aptly expressed it a decade ago: Section 230 ensures that platforms enjoy "power without responsibility."[215]

Unfortunately, platforms' power now includes the ability to ignore the propagation of damaging deep fakes. To be sure, some platforms do not need civil liability exposure to take action against deep-fake generated harms; market pressures and morals are enough. In most cases, however, these forces are insufficient to spur response.

Should Section 230 be amended to extend liability to a wider-range of circumstances? In 2018, lawmakers modified the statute by enacting the Allow States and Victims to Fight Online Sex Trafficking Act ("FOSTA") to address websites' facilitation of sex trafficking.[216] FOSTA added a new exception to

---

207. Citron, *Cyber Civil Rights*, *supra* note 46, at 118.

208. *See* Tim Hwang, Dealing with Disinformation: Evaluating the Case for CDA 230 Amendment (Dec. 17, 2017) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3089442 [https://perma.cc/MD3Q-MR92].

209. Phan v. Pham, 182 Cal. App. 4th 323 (Cal. Ct. App. 2010); Shiamili v. Real Est. Grp. of N.Y., 17 N.Y.3d 281 (N.Y. 2011).

210. Jones v. Dirty World Enter. Recordings, LLC, 755 F.3d 398 (6th Cir. 2014); S.C. v. Dirty World, LLC, No. 11–CV–00392–DW, 2012 WL 3335284 (W.D. Mo. Mar. 12, 2012).

211. Doe v. Backpage.com, LLC, 817 F.3d 12 (1st Cir. 2016).

212. Hinton v. Amazon,com.dedc, LLC, 72 F. Supp. 3d 685, 687 (S.D. Miss. 2014).

213. Cox recently said as much: "I'm afraid . . . the judge-made law has drifted away from the original purpose of the statute." Selyukh, *supra* note 198. In his view, sites that solicit unlawful materials or have a connection to unlawful activity should not enjoy Section 230 immunity. *See id.*

214. *See* Citron, *Cyber Civil Rights*, *supra* note 46, at 118; Mark A. Lemley, *Rationalizing Internet Safe Harbors*, 6 J. TELECOMM. & HIGH TECH. L. 101 (2007); Doug Lichtman & Eric Posner, *Holding Internet Service Providers Accountable*, 14 SUP. CT. ECON. REV. 221 (2006).

215. Rebecca Tushnet, *Power Without Responsibility: Intermediaries and the First Amendment*, 76 GEO. WASH. L. REV. 986 (2008).

216. *See* Danielle Citron & Quinta Jurecic, *FOSTA: The New Anti-Sex-Trafficking Legislation May Not End the Internet, But It's Not a Good Law Either*, LAWFARE (Mar. 28, 2018),

Section 230 immunity, similar to the provision preserving the ability to sue for intellectual property claims. Now, plaintiffs, including state attorneys general, acting on behalf of victims, may avoid Section 230 immunity when suing platforms for knowingly assisting, supporting, or facilitating sex trafficking offenses.

FOSTA did not become law without controversy. Some decried the erosion of Section 230 over concerns that greater liability exposure for online platforms would result in a decrease in outlets, and more self-censorship by those remaining.[217] Others criticized FOSTA's language as indeterminate, potentially resulting in less filtering rather than more.[218] On the other hand, the FOSTA debate also raises the question whether Congress instead erred by not going far enough in carving out exceptions to Section 230 immunity.

Section 230 should be amended to allow a limited degree of platform liability relating to deep fakes. [219] Building on prior work in which one of us (Citron) proposed a similar change in an article co-authored with Benjamin Wittes, we propose that Section 230(c)(1) protections to platforms be conditional rather than automatic.[220] To qualify, an entity must demonstrate that it has taken "reasonable steps" to ensure that its platform is not being used for illegal ends. Platforms that meet this relatively-undemanding requirement will continue to enjoy the protections of Section 230, but others will not and hence may be treated as a publisher of user-generated content that they host.[221]

To be sure, such an amendment would raise hard questions regarding the metes and bounds of reasonableness. The scope of the duty would need to track salient differences among online entities. For example, "ISPs and social networks with millions of postings a day cannot plausibly respond to complaints of abuse immediately, let alone within a day or two,"[222] yet "they may be able to deploy technologies to detect content previously deemed unlawful."[223] Inevitably, the "duty of care will evolve as technology improves."[224]

This proposed amendment would be useful as a means to incentivize platforms to take reasonable steps to minimize the most-serious harms that might follow from user-posted or user-distributed deep fakes. If the reasonably

---

https://www.lawfareblog.com/fosta-new-anti-sex-trafficking-legislation-may-not-end-Internet-its-not-good-law-either [https://perma.cc/2W8X-2KE9] [hereinafter Cintron & Jurecic, *FOSTA*].

217.   *See* CITRON & JURECIC, PLATFORM JUSTICE, *supra* note 46, at 7 (cataloguing the arguments against FOSTA, including the fact that FOSTA raises the moderator's dilemma that animated the adoption of Section 230 and the risk—borne out—that sites will over-filter content related to sex in any way).

218.   *See, e.g.,* Citron & Jurecic, *FOSTA*, *supra* note 216 (arguing that FOSTA is both too narrow and too broad).

219.   Citron & Wittes, *supra* note 46, at 419.

220.   *Id.*

221.   *Id.*

222.   Citron, *Section 230's Challenge*, *supra* note 197.

223.   *Id.*

224.   *Id.*

available technical and other means for detection and removal of harmful fakes are limited, so too will be the obligation on the part of the platform.225 But as those means improve, so would the incentive to use them.226

We recognize that this proposal runs risks, beyond the usual challenges associated with common law development of a novel standard of care. For example, opening the door to liability may over-deter platforms that are uncertain about the standard of care (and fearful of runaway juries imposing massive damages). This might drive sites to shutter (or to never emerge), and it might cause undue private censorship at the sites that remain. Free expression, innovation, and commerce all would suffer, on this view.

To ameliorate these concerns, this proposal can be cabined along several dimensions. First, the amendment to Section 230 could include a sunset provision paired with data-gathering requirements that would empower Congress to make an informed decision on renewal.227 Data-gathering should include the type and frequency of content removed by platforms as well as the extent to which platforms use automation to filter or block certain types of content. This would permit Congress to assess whether the law was resulting in overbroad private censorship, and acting as a Heckler's veto. Second, the amendment could include carefully tailored damages caps. Third, the amendment could be paired with a federal anti-SLAAP provision, which would deter frivolous lawsuits designed to silence protected speech. Last, the amendment could include an exhaustion-of-remedies provision pursuant to which plaintiffs, as a precondition to suit, must first provide notice to the platform regarding the allegedly improper content. The platform would have a specified window of time to examine and respond to the objection.

In sum, a reasonably calibrated standard of care combined with safeguards could reduce opportunities for abuses without interfering unduly with the further development of a vibrant Internet. It would also avoid unintentionally turning innocent platforms into involuntary insurers for those injured through their sites. Approaching the problem with the goal of setting an appropriate standard more readily allows differentiation between kinds of online actors, and a separate rule for websites designed to facilitate illegality in contrast to large ISPs linking millions to the Internet. That said, features used to control the scope of platform

---

225. What comes to mind is Facebook's effort to use hashing technology to detect and remove nonconsensual pornography that has been banned as terms-of-service violations. Citron, *Sexual Privacy*, *supra* note 93, at 1955–58. One of us (Citron) serves on a small task force advising Facebook about the use of screening tools to address the problem of nonconsensually posted intimate images.

226. Current screening technology is far more effective against some kinds of abusive material than others; progress may produce cost-effective means of defeating other attacks. With 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.

227. We see an example of that approach at several points in the history of the "Section 702" surveillance program. Caroline Lynch, *The Virtue of Sunsets?*, LAWFARE (Feb. 28, 2017), https://www.lawfareblog.com/virtue-sunsets [https://perma.cc/5FNL-495P].

liability are only a partial solution to the deep-fakes challenge. Other policy responses will be necessary.

3. *Specific Categories of Criminal Liability*

Civil liability is not the only means through which the legal system can discourage the creation and distribution of harmful deep fakes. Criminal liability is another possibility. Can it close some of the gaps identified above?

Only to a limited extent. The criminal liability model in theory does have the capacity to overcome some of the most significant limits on the civil liability model. Being judgment proof might spare someone from fear of civil suit, for example, but it is no protection from being sent to prison and bearing the other consequences of criminal conviction.[228] And whereas the identification and service of process on the creator or distributor of a harmful deep fake often will be beyond the practical reach of would-be private plaintiffs, law enforcement entities have greater investigative capacities (in addition to the ability to seek extradition). It is far from clear, though, that these notional advantages can be brought to bear effectively in practice.

To some extent, the capacity of criminal law is a question of setting law enforcement priorities and allocating resources accordingly. So far, law enforcement's track record is not promising. Notwithstanding notable exceptions, law enforcement, on the whole, has had a lackluster response to online abuse. In particular, state and local law enforcement agencies often fail to pursue cyberstalking complaints adequately because they lack training in the relevant laws and in the investigative techniques necessary to track down online abusers (federal prosecutors—including especially DOJ's Computer Crimes and Intellectual Property Section—have a much stronger record, but their capabilities do not scale easily).[229] Although a wide range of deep fakes might warrant criminal charges, only the most extreme cases are likely to attract the attention of law enforcement.

Apart from questions of investigative and prosecutorial will, the prospects for criminal liability also depend on the scope of criminal laws themselves. To what extent do existing laws actually cover deep fakes, and to what extent might new ones do so?

A number of current criminal statutes—concerning cyber stalking, impersonation, and defamation—are potentially relevant. Posting deep fakes in connection with the targeting of individuals, for example, might violate the federal cyberstalking laws, 18 U.S.C. § 2261A, or analogous state statutes. Under federal law, it is a felony to use any "interactive computer service or electronic

---

228.   CITRON, HATE CRIMES IN CYBERSPACE, *supra* note 40, at 123.

229.   *Id.* at 144. Assistant US Attorney Mona Sedky is a shining example. *See The Lawfare Podcast: Mona Sedky on Prosecuting Sextortion*, LAWFARE (June 25, 2016), https://www.lawfareblog.com/lawfare-podcast-mona-sedky-prosecutingsextortion [http://perma.cc/262G-KSLV].

communication service" to "intimidate"230 a person in ways "reasonably expected to cause substantial emotional distress . . . ."231 This reflects the fact that, even when cyberstalking victims do not fear bodily harm, "their lives are totally disrupted . . . in the most insidious and frightening ways."232 Defendants can be punished for up to five years in prison and fined up to $250,000, with additional sentencing requirements for repeat offenders and for defendants whose offense violates a restraining order.233 Some deep fakes will fit this bill.

Impersonation crimes may be applicable as well. Several states make it a crime, for example, to knowingly and credibly impersonate another person online with intent to "harm[], intimidat[e], threaten[], or defraud[]" that person.234 And while the "harm, intimidate, threaten" portion of such statutes to some extent tracks the cyberstalking statute described above, its extension to "fraud" opens the door to a wider, though uncertain, range of potential applications. In certain jurisdictions, creators of deep fakes could also face charges for criminal defamation if they posted videos knowing that they were fake or if they were reckless as to their truth or falsity.235 Similarly, using someone's face in a violent deep-fake sex video might support charges for both impersonation and defamation if the defendant intended to terrorize or harm the person and knew the video was fake.

---

230.   18 U.S.C. § 2261A(2) (2012).

231.   18 U.S.C. § 2261A(1)(B). The federal cyberstalking statute has state analogues in a significant number of states, though some state cyberstalking statutes are limited to online abuse sent directly to victims. CITRON, HATE CRIMES IN CYBERSPACE, *supra* note 40, at 124.

232.   *Reauthorization of the Violence Against Women Act: Before the S. Comm. on the Judiciary*, 109th Congress 28 (2005) (statement of Mary Lou Leary, Executive Director of the National Center for Victims of Crime).

233.   18 U.S.C. § 2261A(2).

234.   CAL. PENAL CODE § 528.5 (West 2009); *see also* HAW. REV. STAT. ANN. § 711-1106.6 (2019); LA. REV. STAT. § 14:73.10 (2019); MISS. CODE ANN. § 97-45-33 (2019); N.Y. PENAL LAW § 190.25 (2019); R.I. GEN. LAWS § 11-52-7.1 (2019); TEX. PENAL CODE § 33.07 (2019). The Texas impersonation statute withstood facial challenge in *Ex parte Bradshaw*, 501 S.W.3d 665, 674 (Tex. App. 2016) (explaining that the conduct regulated by the statute is "the act of assuming another person's identity, without that person's consent, and with intent to harm, defraud, intimidate, or threaten . . . by creating a webpage or posting . . ."). Arizona tried to pass a similar law, but the bill failed in the legislature. *See* H.B. 2489, 53 Leg., 1st Sess. (Ariz. 2017). It is a federal crime to impersonate a federal official, though its application may be limited to circumstances in which the defendant intends to defraud others of something of value. 18 U.S.C. § 912 ("Whoever falsely assumes or pretends to be an officer or employee acting under the authority of the United States or any department agency or officer thereof, and acts as such . . . shall be fined under this title or imprisoned."). *Compare* United States v. Gayle, 967 F.2d 483 (11th Cir. 1992) (establishing that an indictment under Sec. 912 did not need to allege an intent to defraud, because such intent could be gathered from the alleged facts), *with* United States v. Pollard, 486 F.2d 190 (5th Cir. 1973) (establishing that failure to allege the intent to defraud is a fatal defect in an indictment under Sec. 912). *See also* United States v. Jones, 16-cr-0553 (AJN), 2018 U.S. Dist. LEXIS 31703 (S.D.N.Y. Feb 2, 2018) (explaining that indictment under § 912 does not include the element to defraud as part of the offense). The 1948 changes to § 912 specifically dropped the words "intent to defraud," yet the Fifth Circuit is the only circuit that still reads the statute to include as an element the intent to defraud.

235.   *See* Eugene Volokh, *One-to-One Speech Vs. One-to-Many Speech, Criminal Harassment Laws, and "Cyberstalking"*, 107 NW. U. L. REV. 731 (2013).

The foregoing examples concern harm to specific individuals, but some harms flowing from deep fakes will be distributed broadly across society. A pernicious example of the latter is a deep fake calculated to spur an audience to violence. Some platforms ban content calling for violence, but not all do.236 Could the creator of such a deep fake be prosecuted under a statute like 18 U.S.C. § 2101, which criminalizes the use of facilities of interstate commerce, such as the Internet, with intent to incite a riot? Incitement charges must comport with the First Amendment constraints identified in *Brandenburg*, including that the speech in question be likely to produce imminent lawless action.237 This leaves many deep fakes beyond the law's reach even though they may have played a role in violence.

Can criminal law be helpful in limiting harms from deep fakes in the particularly sensitive context of elections? Although lies have long plagued the democratic process, deep fakes present a troubling development. Some states have criminalized the intentional use of lies to impact elections.238 These experiments have run into constitutional hurdles, however.

Free speech scholar Helen Norton explains that while political candidates' lies "pose . . . harms to their listeners . . . and may also . . . undermine public confidence in the integrity of the political process," laws forbidding such lies "threaten significant First Amendment harms because they regulate expression in a context in which we especially fear government overreaching and partisan abuse."239 As the Court underscored in *Brown v. Hartlage*,240 the "State's fear that voters might make an ill-advised choice does not provide the State with a compelling justification for limiting speech."241 Not surprisingly, courts therefore have struck down periodic attempts to ban election-related lies.242 The entry of deep fakes into the mix may not change that result. As explored above,

---

236. YouTube, for example, barred incitement in 2008. *See* Peter Whoriskey, *YouTube Bans Videos That Incite Violence,* WASH. POST (Sept. 12, 2008), http://www.washingtonpost.com/wp-dyn/content/article/2008/09/11/AR2008091103447.html [https://perma.cc/YVR5-JGXV].

237. Multiple states prescribe criminal penalties for those who engage in similar conduct. *See, e.g.*, CAL. PENAL CODE § 404.6 (2019); FLA. STAT. ANN. § 870.01 (2019); MONT. CODE ANN. § 45-8-105 (2019); VA. CODE ANN. § 18.2-408 (2019). For an excellent overview of crimes of incitement in the digital age and the associated issues, see Margot E. Kaminski, *Incitement to Riot in the Age of Flash Mobs*, 81 U. CIN. L. REV. 1 (2012).

238. *See* Nat Stern, *Judicial Candidates' Right to Lie*, 77 MD. L. REV. 774 (2018).

239. Richard L. Hasen, *A Constitutional Right to Lie in Campaigns and Elections?*, 74 MONT. L. REV. 53, 69 (2013) ("[T]o survive constitutional review, any false campaign speech law would have to be narrow, targeted only at false speech made with actual malice . . . ."); Helen Norton, *Lies and the Constitution*, 2012 SUP. CT. REV. 161, 199 (2012).

240. 456 U.S. 45, 46 (1982).

241. *Id.* at 60.

242. *See, e.g.,* Susan B. Anthony List v. Driehaus, 814 F.3d 466 (6th Cir. 2016) (striking down an Ohio election-lies law as a content-based restriction of "core political speech" that lacked sufficient tailoring); 281 Care Comm. v. Arneson, 766 F.3d 774, 785 (8th Cir. 2014) ("[N]o amount of narrow tailoring succeeds because [Minnesota's political false-statements law] is not necessary, is simultaneously overbroad and underinclusive, and is not the least restrictive means of achieving any stated goal.").

however, criminal laws banning the impersonation of government officials or candidates for office may overcome constitutional challenge.[243]

Ultimately, criminal liability is not likely to be a particularly effective tool against deep fakes that pertain to elections. The most capable actors with motive and means to deploy deep fakes in a high-impact manner in an election setting will include the intelligence services of foreign governments engaging in such activity as a form of covert action, as we saw with Russia in relation to the American election of 2016. The prospect of a criminal prosecution in the United States will mean little to foreign government agents involved in such activity so long as they are not likely to end up in US custody (though it might mean something more to private actors through whom those agencies sometimes choose to act, at least if they intend to travel abroad).[244]

C. *Administrative Agency Solutions*

The foregoing analysis suggests that prosecutors and private plaintiffs can and likely will play an important role in curbing harms from deep fakes, but also that this role has significant limitations. We therefore turn to consider the potential contributions of other actors, starting with administrative agencies.

Generally speaking, agencies can advance public policy goals through rulemaking, adjudication, or both.[245] Agencies do not enjoy plenary jurisdiction to use these tools in relation to any subject they wish. Typically, their field of operation is defined—with varying degrees of specificity—by statute. And thus we might begin by asking which agencies have the most plausible grounds for addressing deep fakes.

At the federal level, three candidates stand out: the Federal Trade Commission ("FTC"), the Federal Communications Commission ("FCC"), and the Federal Election Commission ("FEC"). On close inspection, however, their potential roles appear quite limited.

1. *The FTC*

Consider the Federal Trade Commission and its charge to regulate and litigate in an effort to minimize deceptive or unfair commercial acts and practices.[246] For that matter, consider the full range of state actors (often a state's

---

243.   *See supra* notes 170-172 and accompanying text. For a thoughtful exploration of why deep fakes created and used in election context should be understood as proscribable fraud, see Green, *supra* note 108.

244.   On the use of private actors by state agencies in the context of hacking, see TIM MAURER, CYBER MERCENARIES: THE STATE, HACKERS, AND POWER (2018). For an example of successful prosecution of such private actors, see United States v. Baratov, No. 3:17-CR-103 VC, 2018 WL 1978898 (N.D. Cal. Apr. 17, 2018) (five-year sentence for Canadian national who acted as a contractor involved in a hacking campaign directed by Russia's FSB against companies including Yahoo!).

245.   *See* Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1278 (2008).

246.   5 U.S.C. § 45(b) (2012).

Attorney General's Office) that play a similar role. Bearing that charge in mind, can these entities intervene in the deep fake context?

A review of current areas of FTC activity suggests limited possibilities. Most deep fakes will not take the form of advertising, but some will. That subset will implicate the FTC's role in protecting consumers from fraudulent advertising relating to "food, drugs, devices, services, or cosmetics."[247] Some deep fakes will be in the nature of satire or parody, without intent or even effect of misleading consumers into believing a particular person (a celebrity or some other public figure) is endorsing the product or service in question. That line will be crossed in some instances, however. If such a case involves a public figure who is aware of the fraud and both inclined to and capable of suing on their own behalf for misappropriation of likeness, there is no need for the FTC or a state agency to become involved. Those conditions will not always be met, though, especially when the deep-fake element involves a fraudulent depiction of something other than a specific person's words or deeds; there would be no obvious private plaintiff. The FTC and state attorneys general (state AGs) can play an important role in that setting.

Beyond deceptive advertising, the FTC has authority to investigate unfair and deceptive commercial acts and practices under Section 5 of the Federal Trade Commission Act.[248] Much like Section 5 of the Federal Trade Commission Act, state UDAP laws (enforced by state AGs) prohibit deceptive commercial acts and practices and unfair trade acts and practices whose costs exceed their benefits.[249] UDAP laws empower attorneys general to seek civil penalties, injunctive relief, and attorneys' fees and costs.[250]

Acting in that capacity, for example, the FTC previously investigated and reached a settlement with Facebook regarding its treatment of user data—and is now doing so again in the aftermath of public furor over the Cambridge Analytica debacle.[251] In response to the problem of fake news in general and deep-fake news in particular, the FTC might contemplate asserting a role under

---

247. 5 U.S.C. § 52(a)(1)–(2) (2012).

248. *See* 15 U.S.C. § 45. For the crucial role that the FTC has played in the development of privacy policy, see CHRIS JAY HOOFNAGLE, FEDERAL TRADE COMMISSION PRIVACY LAW AND POLICY (2016); Woodrow Hartzog & Daniel J. Solove, *The Scope and Potential of FTC Data Protection*, 83 GEO. WASH. L. REV. 2230 (2015); Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583 (2014).

249. *See generally* Danielle Keats Citron, *The Privacy Policymaking of State Attorneys General*, 92 NOTRE DAME L. REV. 747, 755–57 (2016).

250. *See, e.g.*, California Unfair Business Act, CAL. BUS. & PROF. CODE § 17206 (West 2016) (imposing $ 2,500 per violation); Illinois Consumer Fraud Act, 815 ILL. COMP. STAT. ANN. 505/7 (West 2016) (allowing civil penalty of $50,000 per unlawful act); *see also* Steven J. Cole, *State Enforcement Efforts Directed Against Unfair or Deceptive Practices,* 56 ANTITRUST L.J. 125, 128 (1987) (explaining that in states like Maryland the "consumer protection authority resides wholly within the Attorney General's Office").

251. Louise Matsakis, *The FTC is Officially Investigating Facebook's Data Practices*, WIRED (Mar. 26, 2018), https://www.wired.com/story/ftc-facebook-data-privacy-investigation [https://perma.cc/GJX8-LQ27].

the rubric of "unfair or deceptive acts or practices in or affecting commerce."252 Any such efforts would face several obstacles, however. First, Section 230 of the Communications Decency Act as currently written would shield platforms at least to some extent from liability for publishing users' deep fakes. Second, it is not clear this would be a proper interpretation of the FTC's jurisdiction. Professor David Vladeck, formerly head of the FTC's Bureau of Consumer Protection, has expressed doubt about the FTC's jurisdiction to regulate sites purveying fake news.253 Vladeck argues, "[f]ake news stories that get circulated or planted or tweeted around are not trying to induce someone to purchase a product; they're trying to induce someone to believe an idea."254 Finally, the prospect of a government entity attempting to distinguish real news from fake news—and suppressing the latter—raises the First Amendment concerns described above in relation to election-lies laws.

Might a different agency at least have a stronger jurisdictional claim to become involved in some settings? This brings us to the Federal Communications Commission.

2. *The FCC*

If any regulatory agency is to play a role policing against harms from deep fakes circulating online, the FCC at first blush might seem a natural fit. It has a long tradition of regulating the communications of broadcasters, and many have observed that the major social media platforms of the twenty-first century occupy a place in our information ecosystem similar to the central role that radio and television broadcasters enjoyed in the twentieth century.255 Similar thinking led the FCC in 2015 to break new ground by reclassifying Internet service providers as a "telecommunications service" rather than an "information service," thus opening the door to more extensive regulation.256 Amidst intense controversy, however, the FCC in late 2017 reversed course on this position on ISPs,257 and in any event never asserted that so-called "edge providers" like Facebook also should be brought under the "telecommunications service" umbrella.258

---

252.   Federal Trade Commission Act, 15 U.S.C. § 45(a)(1) (2012); *see* Callum Borchers, *How the Federal Trade Commission Could (Maybe) Crack Down on Fake News*, WASH. POST (Jan. 30, 2017),   https://www.washingtonpost.com/news/the-fix/wp/2017/01/30/how-the-federal-trade-commission-could-maybe-crack-down-on-fake-news/?utm_term=.4ef8ece1baec [https://perma.cc/L2XD-T445].

253.   *Id.*

254.   *Id.*

255.   *See* TIM WU, THE MASTER SWITCH: THE RISE AND FALL OF INFORMATION EMPIRES (2010).

256.   *See* Protecting and Promoting the Open Internet, 80 Fed. Reg. 19,737 (F.C.C. Apr. 13, 2015) (declaratory ruling).

257.   Restoring Internet Freedom, FCC 17-166 (2018).

258.   Consumer Watchdog Petition for Rulemaking to Require Edge Providers to Honor 'Do Not Track' Requests, DA 15-1266 (2015).

As things stand, the FCC appears to lack jurisdiction (not to mention interest) over content circulated via social media. However, concern over fake news, incitement, radicalization, or any number of other hot-button issues might at some point tip the scales either for the FCC to reinterpret its own authority or for Congress to intervene. For the moment, however, this pathway appears closed, leaving the FCC's role in relation to deep fakes limited to potential efforts to deter their appearance on radio or television.

3. *The FEC*

A third federal agency with a plausible stake in the topic of deep fakes is the Federal Election Commission. Plainly, its jurisdiction would touch upon deep fakes only as they relate to elections—a narrow, but important, subfield. Whether and how the FEC might act in relation to deep fakes even in that setting, however, is unclear.

The FEC regulates campaign speech, but not in ways that would speak directly to the deep-fake scenario. In particular, the FEC does not purport to regulate the truth of campaign-related statements, nor is it likely to assert or receive such jurisdiction anytime soon for all the reasons discussed above in relation to the First Amendment obstacles, practical difficulty, and political sensitivity of such an enterprise. Instead, its central focus is financing, and the main thrust of its regulatory efforts relating to speech is to increase transparency regarding sponsorship and funding for political advertising.[259]

There might be room for a regulatory approach that requires deep fake creators to disclose the fact that the video or audio is a fake.[260] The Court has upheld campaign speech regulations requiring truthful disclosure of the source of the communication.[261] And for good reason—listeners depend upon the source of speech to make decisions at the ballot box.[262]

Such an approach could have at least some positive impact on deep fakes in the electoral setting. For outlets within the FEC's jurisdiction, transparency obligations create elements of attribution and accountability for content creators that might, to some extent, deter resort to deep fakes in advertising. But note that major online social media platforms are not, currently, subject to FEC jurisdiction in this context: Facebook, Google, and other online advertising platforms have long-resisted imposition of the FEC's disclosure rules, often

---

259.   For an interesting proposal for new regulations that the FEC might fruitfully pursue in this vein with respect to the general problem of misleading campaign advertising, see Abby K. Wood & Ann M. Ravel, *Fool Me Once: Regulating "Fake News" and Other Online Advertising*, 91 S. CAL. L. REV. 1223 (2018).

260.   Blitz, Deep Fakes and Other Non-Testimonial Falsehoods, *supra* note 165.

261.   Norton, *Lies to Manipulate*, *supra* note 169, at 165–67.

262.   Michael S. Kang, *Democratizing Direct Democracy: Restoring Voter Competence Through Heuristic Cues and 'Disclosure Plus,'* 50 UCLA L. REV. 1141, 1158–59 (2003); Helen Norton, *Secrets, Lies, and Disclosure*, 27 J.L. & POL. 641, 644 (2012).

citing the practical difficulties that would follow for small screens displaying even smaller ads.

In the wake of the 2016 election, the FEC faces pressure to extend its reach to these platforms nonetheless, so that caveat might drop out at some point.263 Even so, this certainly would not resolve the threat to elections posed by deep fakes.

FEC regulation surely would not eliminate deep fakes' threat to elections. Some amount of fraudulent posting no doubt would continue simply because enforcement systems will not be perfect, and also because not all content about someone who is a candidate will be framed in ways that would appear to count as advertising. Deep fakes in particular are likely to take the form of just raw video or audio of some event that occurred, by no means necessarily embedded within any larger narrative or framing content. The FEC's disclosure rules in any event are candidate specific, and do not encompass generalized "issue ads" that express views on a topic but do not single out particular candidates.

D.   *Coercive Responses*

The utility of civil suits, criminal prosecution, and regulatory actions will be limited when the source of the fake is a foreign entity that may lie beyond the reach of American judicial process (though it is not non-existent, as we have seen from time to time in the context of cybersecurity).264 Nevertheless, it is important to recall that the Government possesses other instruments that it can bring to bear in such contexts in order to impose significant costs on the perpetrators. We provide a brief discussion of three such scenarios here.

1.   *Military Responses*

There is no doubt that deep fakes will play a role in future armed conflicts. Information operations of various kinds have long been an important aspect of warfare, as the contending parties attempt to influence the beliefs, will, and passions of a wide range of audiences (opposing forces and their commanders,

---

263.   Google in 2006 obtained an exemption from disclosure obligations based on the practical argument that its online ad spaces were too small to accommodate the words. In the spring of 2018 the FEC began the process of changing this approach. *See* Alex Thompson, *The FEC Took a Tiny Step to Regulate Online Political Ads, But Not in Time for 2018 Elections*, VICE NEWS (Mar. 15, 2018), https://news.vice.com/en_us/article/neq88q/the-fec-took-a-tiny-step-to-regulate-online-political-ads-but-not-in-time-for-2018-elections [https://perma.cc/E7QB-NXAW].

264.   For example, foreign nationals at times have been extradited to the United States to face criminal charges relating to hacking. *See* Press Release, U.S. Attorney's Office for the Southern District of New York, "Manhattan U.S. Attorney Announces Extradition Of Alleged Russian Hacker Responsible For Massive Network Intrusions At U.S. Financial Institutions, Brokerage Firms, A Major News Publication, And Other Companies" (Sept. 7, 2018), https://www.justice.gov/usao-sdny/pr/manhattan-us-attorney-announces-extradition-alleged-russian-hacker-responsible-massive [https://perma.cc/2A36-LXDD].

opposing politicians and electorates, local populations, allies, and so forth).[265] Such effects are sought at every level from the tactical to the strategic, and with an eye towards effects ranging from the immediate to the long-term.

Deep-fake capacity will be useful in all such settings. Insurgents, for example, might inflame local opinion against US or allied forces by depicting those forces burning a Quran or killing a civilian. If deployed deftly enough, such fraud might also be used to advance a "lawfare" strategy, leveraging the good intentions of journalists and NGOs to generate distracting or even debilitating legal, political, and diplomatic friction. Insurgents also might deploy the technology to make their own leaders or personnel appear more admirable or brave than otherwise might be possible, to create the false impression that they were in a particular location at a particular time, or even to make it seem that a particular leader is still alive and free rather than dead or captured. The US military, for its part, might use deep fakes to undermine the credibility of an insurgent leader by making it appear that the person is secretly cooperating with the United States or engaging in immoral or otherwise hypocritical behavior. If the technology is robust enough, and deployed deftly enough, the opportunities for mischief—deadly mischief, in some cases—will be plentiful on both sides.

If and when adversaries of the United States do use deep fakes in connection with an armed conflict, the options for a military response would be no different than would be the case for any form of enemy information operation. This might entail penetration of the adversary's computer networks, for purposes of both intelligence gathering, making it easier to prepare for or respond to a deep fake, and disruption operations, degrading or destroying the adversary's capacity to produce them in the first place. It might entail a kinetic strike on facilities or individuals involved in the deep fake production process, subject of course to the law of armed conflict rules governing distinction, proportionality, and so forth.[266] And it might entail the capture and detention of enemy personnel or supporters involved in such work.

---

265. The US military defines "information operations," as the use of any and all information-related capabilities during the course of military operations in order "to influence, disrupt, corrupt, or usurp adversarial human and automated decision-making while protecting our own." CHAIRMAN OF THE JOINT CHIEFS OF STAFF, JOINT PUBLICATION 3-13: PSYCHOLOGICAL OPERATIONS VI-5 (2010). Separately, it defines "psychological operations" as "planned operations to convey selected information and indicators to foreign audiences to influence their emotions, motives, objective reasoning, and ultimately the behavior of foreign governments, organizations, groups, and individuals" in a manner "favorable to the originator's objectives." *Id.* at GL-8. Until 2010, these activities were known as psychological operations, or psyops. In 2017, the Army re-adopted the psyops name. *See MISO Name Change—Back to Psychological Operations (PSYOP)*, SOF NEWS (Nov. 8, 2017), http://www.sof.news/io/miso-name-change [https://perma.cc/79VX-XN8B].

266. The possibility of targeting a person based solely on involvement in production of a deep-fake video supporting the enemy—as opposed to targeting them based on status as a combatant—would raise serious issues under the principle of distinction. Assuming, again, that the prospective target is best categorized as a civilian, he or she would be targetable only while directly participating in hostilities. Debates abound regarding the scope of direct participation, but most scenarios involving creation of media would appear to be indirect in nature. One can imagine a special case involving especially

The situation becomes more complicated insofar as the individuals or servers involved in creating deep fakes relating to an armed conflict are not actually located in theater. If either reside in third countries, the freedom of action for a military response of any kind may be sharply circumscribed both by policy and by legal considerations. This is a familiar challenge for the military in relation to non-deep-fake online propaganda activity conducted by and for the Islamic State using servers outside the Syria/Iraq theater, and the manner in which it would play out would be no different (for better or worse) if one introduces deep-fake technology to the mix.

2. Covert Action

Covert action might be used as a response to a foreign government's use of deep fakes. "Covert action" refers to government-sponsored activity that is meant to impact events overseas without the US government's role being apparent or acknowledged.[267] That is a capacious definition, encompassing a wide-range of potential activities. Propaganda and other information operations, for example, can be and frequently are conducted as covert actions. And certainly we can expect to see the intelligence services of many countries making use of deep-fake technologies in that context in the future (the Russian covert action campaign that targeted the American election in 2016 was significant even without the aid of deep fakes, but one can certainly expect to see deep fakes used in such settings in the future). The point of mentioning covert action here is not to repeat the claim that states will use deep fakes on an unacknowledged basis in the future. Instead, the point is to underscore that the US government has the option of turning to covert action *in response* to a foreign government's use of deep fakes.

What, in particular, might this entail? First, it could be the basis for degrading or destroying the technical capacity of a foreign actor to produce deep fakes (for example, through a computer network operation designed to make subtle changes to a GAN). The military options described above also included such technical means, but covert action offers advantages over the military alternative. Most notably, covert action does not require any predicate circumstance of armed conflict; presidents may resort to it when they wish. Moreover, because covert action is not publicly acknowledged, the diplomatic and political friction that might otherwise make a particular action unattractive is reduced in comparison to overt alternatives (although not necessarily eliminated, for the activity may later become public). Further, covert action may be a particularly attractive option where the activity in question might violate international law. The statutory framework governing covert action requires

---

inflammatory deep fakes designed to cause an immediate violent response, though even there hard questions would arise about the likely gap in time between creation of such a video and its actual deployment.

267.   *See* 50 U.S.C. § 3093(e) (2012).

compliance with the Constitution and statutes of the United States, but it is conspicuously silent about compliance with international law. Many have speculated that this is construed within the government as domestic-law justification for activities that violate international law.[268]

Covert action can take any number of other forms. Rather than directly disrupting a foreign target's capacity to produce deep fakes, for example, covert means might be used in a wide variety of ways to impose costs on the person, organization, or government at issue. Covert action, in other words, can be used to deter or punish foreign actors that employ deep fakes in ways harmful to the United States.[269]

Covert-action tools are not the only options the US government has with respect to imposing costs on foreign individuals or entities who may make harmful use of deep fakes. We turn now to a brief discussion of a leading example of an overt tool that can serve this same purpose quite effectively.

3. Sanctions

The economic might the United States developed over the past half-century has given the US Government considerable leverage over foreign governments, entities, and individuals. Congress, in turn, has empowered the executive branch to move quickly and largely at the president's discretion when it wishes to exploit that leverage to advance certain interests. Most notably for present purposes, the International Emergency Economic Powers Act ("IEEPA") establishes a framework for the executive branch to issue economic sanctions backed by criminal penalties.[270]

In order to bring this power to bear, IEEPA requires that the president first issue a public proclamation of a "national emergency" relating to an "unusual and extraordinary threat, which has its source in whole or substantial part outside the United States."[271] In order to deploy IEEPA sanctions as an overt response to foreign use of deep fakes, therefore, there needs to be either a relevant existing national-emergency proclamation or else plausible grounds for issuing a new one towards that end.

There is no current national-emergency proclamation that would apply generally to the problem of deep fakes. There are more than two-dozen currently active states of national emergency, as of the summer of 2018.[272] Most have little

---

268.   *See* Robert M. Chesney, *Computer Network Operations and U.S. Domestic Law: An Overview*, 89 INT'L L. STUD. 218, 230–32 (2013).

269.   Covert action cannot have this deterrent effect, however, if the targeted person or entity is unaware that the United States imposed those costs, and that it did so for a particular reason. This is a tricky (but by no means insurmountable) obstacle where the sponsoring role of the United States is not meant to be acknowledged publicly.

270.   *See* 50 U.S.C. ch. 35 (2012).

271.   50 U.S.C. § 1701(a) (2012).

272.   *See* Ryan Struyk, *Here are the 28 Active National Emergencies. Trump Won't Be Adding the Opioid Crisis to the List*, CNN POL. (Aug. 15, 2017),

possible relevance, but some relate broadly to particular threat actors or regions, and a deep-fake scenario conceivably might arise in ways that both implicate those actors or regions and involve actors not already subject to sanctions.

A particularly important question under this heading is whether any of these existing authorities would apply to a foreign entity employing deep fakes to impact American elections. The answer appears to be yes, though the matter is complicated.

In April 2015, President Obama's Executive Order 13964 proclaimed a national emergency with respect to "malicious cyber-enabled activities originating from, or directed by persons located . . . outside the United States."[273] Then, in the aftermath of the 2016 election, Obama amended the order, expanding the prohibition to forbid foreign entities from using cyber-enabled means to "tamper[] with, alter[], or caus[e] a misappropriation of information with the purpose or effect of interfering with or undermining election processes or institutions . . . ."[274] This was designed to allow for IEEPA sanctions against Russian entities that interfered in the 2016 election through means that included the DNC hack.

President Obama immediately used the authority to sanction Russia's FSB, GRU, and various other individuals and entities.[275] But could the same be done to a foreign entity that had not engaged in hacking, and instead focused entirely on using social media platforms to propagate false information in ways meant to impact American politics?[276]

To the surprise of some observers, the Trump administration provided at least a degree of support for the broader interpretation in March 2018 when it issued sanctions against Russia's Internet Research Agency (IRA) under color

https://www.cnn.com/2017/08/12/politics/national-emergencies-trump-opioid/index.html [https://perma.cc/B9BW-PSAR]; *see also* Catherine Padhi, *Emergencies Without End: A Primer on Federal States of Emergency*, LAWFARE (Dec. 8, 2017), https://lawfareblog.com/emergencies-without-end-primer-federal-states-emergency [https://perma.cc/FW7X-PG75].

273.   Exec. Order No. 13694, 80 Fed. Reg. 18,077 (Apr. 1, 2015).

274.   Exec. Order No. 13757, 82 Fed. Reg. 1 (Dec. 28, 2016).

275.   *See Issuance of Amended Executive Order 13694; Cyber-Related Sanctions Designation*, U.S. DEP'T TREASURY (Dec. 29, 2016), https://www.treasury.gov/resource-center/sanctions/OFAC-Enforcement/Pages/20161229.aspx [https://perma.cc/7A6G-NUVL].

276.   The Treasury Department has indicated that it will promulgate regulations defining "cyber-enabled activities," and in the meantime has offered a less-formal explanation of its view that emphasizes unauthorized access, yes, but also includes much broader language: "We anticipate that regulations to be promulgated will define 'cyber-enabled' activities to include *any act that is primarily accomplished through or facilitated by computers or other electronic devices*. For purposes of E.O. 13694, malicious cyber-enabled activities include deliberate activities accomplished through unauthorized access to a computer system, including by remote access; circumventing one or more protection measures, including by bypassing a firewall; or compromising the security of hardware or software in the supply chain. These activities are often the means through which the specific harms enumerated in the E.O. are achieved, including compromise to critical infrastructure, denial of service attacks, or massive loss of sensitive information, such as trade secrets and personal financial information." (emphasis added). *OFAC FAQs: Other Sanctions Programs*, U.S. DEP'T TREASURY, https://www.treasury.gov/resource-center/faqs/Sanctions/Pages/faq_other.aspx [https://perma.cc/JPB9-W29J].

of this framework.277 The IRA, infamously, had engaged in extensive efforts to propagate false information into the American political debate. When the Trump administration sanctioned it under color of the cyber executive order, this seemed to endorse the proposition that politically targeted information operations carried out online were enough, even without hacking, to trigger the IEEPA framework. A close read of the Treasury Department's explanation of IRA's inclusion, however, includes just enough reference to "misappropriation of information" and to illegal use of stolen personally identifiable information so as to muddy the precedent.278

Bearing this lingering uncertainty in mind, we recommend promulgation of a new national emergency specifically tailored to attempts by foreign entities to inject false information into America's political dialogue, without any need to show that such efforts at some point happened to involve hacking or any other "cyber-enabled" means. This would eliminate any doubt about the immediate availability of IEEPA-based sanctions. Attempts to employ deep fakes in aid of such efforts would, of course, be encompassed in such a regime.

E.   Market Solutions

We anticipate two types of market-based reactions to the deep-fake threat. First, we expect the private sector to develop and sell services intended to protect customers from at least some forms of deep fake-based harms. Such innovations might build on the array of services that have emerged in recent years in response to customer anxieties about identity theft and the like. Second, we expect at least some social media companies to take steps on their own initiative to police against deep-fake harms on their platforms. They will do this not just because they perceive market advantage in doing so, of course, but also for reasons including policy preferences and, perhaps, concern over what legislative interventions, including amendments to Section 230 of the Communications Decency Act, might occur down the road if they take no action. Both prospects offer benefits, but there are both limits and risks as well.

---

277.   Press Release, U.S. Dep't Treasury, Treasury Sanctions Russian Cyber Actors for Interference with the 2016 U.S. Elections and Malicious Cyber Attacks (Mar. 15, 2018) https://home.treasury.gov/news/press-releases/sm0312 [https://perma.cc/2YRG-68XQ].

278.   *See id.* ("The Internet Research Agency LLC (IRA) tampered with, altered, or caused a misappropriation of information with the purpose or effect of interfering with or undermining election processes and institutions. Specifically, the IRA tampered with or altered information in order to interfere with the 2016 U.S. election. The IRA created and managed a vast number of fake online personas that posed as legitimate U.S. persons to include grassroots organizations, interest groups, and a state political party on social media. Through this activity, the IRA posted thousands of ads that reached millions of people online. The IRA also organized and coordinated political rallies during the run-up to the 2016 election, all while hiding its Russian identity. Further, the IRA unlawfully utilized personally identifiable information from U.S. persons to open financial accounts to help fund IRA operations.").

1. *Immutable Life Logs as an Alibi Service*

Consider a worst-case scenario: a world in which it is cheap and easy to portray people as having done or said things they did not say or do, with inadequate technology to quickly and reliably expose fakes and inadequate law or policy tools to deter and punish them. In that environment, a person who cannot credibly demonstrate their real location, words, and deeds at a given moment will be at greater risk than those who can. Credible alibis will become increasingly valuable as a result; demand for new ways to secure them—for services that ensure that one can disprove a harmful fake—will grow, spurring innovation as companies see a revenue opportunity.

We predict the development of a profitable new service: immutable life logs or authentication trails that make it possible for a victim of a deep fake to produce a certified alibi credibly proving that he or she did not do or say the thing depicted.279

From a technical perspective, such services will be made possible by advances in a variety of technologies including wearable tech; encryption; remote sensing; data compression, transmission, and storage; and blockchain-based record-keeping. That last element will be particularly important, for a vendor hoping to provide such services could not succeed without earning a strong reputation for the immutability and comprehensiveness of its data; otherwise, the service would not have the desired effect when called upon in the face of an otherwise-devastating deep fake.

Providing access to a credible digital alibi would not be enough, however. The vendor also would need to be able to provide quick and effective dissemination of it; the victim alone often will be in a poor position to accomplish that, for the reasons discussed above in Part I. But it is possible that one or a few providers of an immutable life log service can accomplish this to no small degree. The key would be partnerships with a wide array of social media platforms, with arrangements made for those companies to rapidly and reliably coordinate with the provider when a complaint arises regarding possible deep-fake content on their site.

Obviously, not everyone would want such a service even if it could work reasonably effectively as a deep-fake defense mechanism. But some individuals (politicians, celebrities, and others whose fortunes depend to an unusual degree on fragile reputations) will have sufficient fear of suffering irreparable harm from deep fakes that they may be willing to agree to—and pay for—a service that comprehensively tracks and preserves their movements, surrounding visual circumstances, and perhaps in-person and electronic communications; although providers may be reluctant to include audio-recording capacity because some

---

279.   This notion is by no means new. Indeed, Anita Allen presciently discussed this possibility in her work. *See* Anita L. Allen, *Dredging Up the Past: Lifelogging, Memory, Surveillance*, 75 U. CHI. L. REV. 47 (2008).

states criminalize the interception of electronic communications unless all parties to a communication consent to the interception.280

Of course, a subset of such a service—location verification—is available already, thanks to the ubiquity of phones with location tracking features as well as cell-site location records. But it is one thing to have theoretical access to a business record proving that a device (though not necessarily the person associated with it) was in some general location. It would be quite another to have ready and reliable access to proof—perhaps backed by video—that the person was in a very precise location and acting and speaking in particular ways. And if the provider of such a service manages to partner with major platforms in a way that facilitates not just reliable but rapid and efficient verification services, this could be a sizable advantage.

Even so, it may be that few individuals will want to surrender privacy in this way. We think it likely, though, that more than a few organizations will consider requiring use of tracking services by at least some employees at least some of the time. The protective rationale for the service will be a considerable incentive for the organization, but note that this interest might dovetail robustly with distinct managerial interests in deterring or catching employee misfeasance and malfeasance. This is much like the earlier wave of innovation that led to installation of dashboard cameras in police cars and the current wave involving the proliferation of body cameras on the officers themselves.

We urge caution in encouraging the emergence of such services. Whatever the benefits, the social cost (should such services emerge and prove popular) would be profound.

Proliferation of comprehensive life logging would have tremendous spillover impacts on privacy in general. Indeed, it risks what has been called the "unraveling of privacy"281—the outright functional collapse of privacy via social consent despite legal protections intended to preserve it. Scott Peppet has warned that, as more people relinquish their privacy voluntarily, the remainder increasingly risks being subject to the inference that they have something to hide.282 This dynamic might eventually overcome the reluctance of some holdouts. Worse, the holdouts in any event will lose much of their lingering privacy, as they find themselves increasingly surrounded by people engaged in life-logging.

Note the position of power in which this places the suppliers of these services. The scale and nature of the data they would host would be

---

280.   *See* Danielle Keats Citron, *Spying Inc.*, 72 WASH. & LEE L. REV. 1243, 1263 (2015) (explaining that twelve states criminalize the interception of electronic communications unless all parties to the communication consent to the interception); Paul Ohm, *The Rise and Fall of Invasive ISP Surveillance*, 2009 U. ILL. L. REV. 1417, 1486 (2009). So long as one party to communications consent to interception, the remaining state laws—38—and federal law permit the practice.

281.   Scott R. Peppet, *Unraveling Privacy: The Personal Prospectus and the Threat of a Full-Disclosure Future*, 105 NW. U. L. REV. 1153, 1159 (2015).

282.   *Id.* at 1180.

extraordinary, both as to individual clients and more broadly across segments of society or even society as a whole. A given company might commit not to exploit that data for commercial or research purposes, hoping instead to draw revenue solely from customer subscriptions. But the temptation to engage in predictive marketing, or to sell access to the various slices of the data, would be considerable. The company would possess a database of human behavior of unprecedented depth and breadth, after all, or what Paul Ohm has called a "database of ruin."[283] The Cambridge Analytica/Facebook scandal might pale in comparison to the possibilities unleashed by such a database.

The existence of such a database would also raise privacy issues vis-à-vis government investigators. Certainly law enforcement entities would wish to access this rich trove of information in many cases.[284] Whether they could do so without a warrant, however, is unclear at the current time. The Supreme Court's 2018 decision in *Carpenter v. United States* unsettled the so-called "third-party doctrine" (i.e., the rule that the Fourth Amendment does not require a warrant for government access to records held by a third party).[285] While *Carpenter* disclaimed any intent to abandon the third-party doctrine with respect to "conventional surveillance techniques and tools, such as security cameras,"[286] the opinion suggests that a warrant likely would be required in the case of a police search of a database of the kind created by comprehensive life logging. Indeed, a life-logging database would enable precisely the sort of pervasive surveillance of someone's life that triggered the warrant for access to cell-site location data.[287] Congress or state legislatures might directly impose such a requirement by statute. But at any rate, the important point is that—once the right legal process is used—the government's capacity to know all about a suspect would be unrivaled as a historical matter (especially as combined with other existing aggregations of data).

---

283. Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization,* 57 UCLA L. REV. 1701, 1748 (2010).

284. *See* Neil Richards, *The Third Party Doctrine and the Future of the Cloud*, 94 WASH. U. L. REV. 1441, 1444 (2017).

285. Carpenter v. United States, 138 S. Ct. 2206 (2018). For an insightful discussion of the Carpenter decision, see Paul Ohm, *The Many Revolutions of Carpenter*, 32 HARV. J. L. & TECH. 357 (2019).

286. *Carpenter*, 138 S. Ct. at 2220.

287. The *Carpenter* decision follows logically from the opinions articulated in *United States v. Jones*, 565 U.S. 400 (2012), which David Gray and one of us (Citron) argued amounted to the recognition of a right to quantitative privacy. *See* David Gray & Danielle Citron, *The Right to Quantitative Privacy*, 98 MINN. L. REV. 62, 64–65 (2013) (arguing that the Fourth Amendment erects protection against broad and indiscriminate surveillance that is tantamount to a general warrant). Though the third-party doctrine was not actually modified in *United States v. Jones*, five justices in that case expressed doubt about the wisdom of simply applying the third-party doctrine unaltered to circumstances involving novel information technologies that do not necessarily track the premises of the analog age that gave rise to that doctrine and that raise the spectre of a surveillance state. 565 U.S. at 89–92.

Despite helping to identify those guilty of crime and avoid mistaken prosecution of the innocent, this would produce unprecedented opportunities for government authorities to stumble across—and then pursue—*other* misdeeds, and not only those of the original suspect. Society may not be prepared to accept what might then be a sharp increase in the degree of detection and enforcement that would follow. Moreover, the situation also would expose investigators to a considerable amount of information that might not be inculpatory as such, but that might, nonetheless, provide important leverage over the suspect or others. Again, the resulting enhancement of prosecutorial capacity will be welcome in some quarters, but may cause an erosion of commitment to privacy and other values. At the very least, this would deserve careful consideration by policymakers and lawmakers.

Ultimately, a world with widespread life logging of this kind might yield more good than harm, particularly if paired with legislation guarding access to, use of, and security accorded such comprehensive databases. But it might not. For now, our aim is no more and no less than to identify the possibility that the rise of deep fakes will in turn give birth to such a service, and to flag the implications this will have for privacy. Enterprising businesses may seek to meet the pressing demand to counter deep fakes in this way, but it does not follow that society should welcome—or wholly accept—that development. Careful reflection is essential now, before *either* deep fakes *or* responsive services get too far ahead of us.

2. *Speech Policies of Platforms*

Our last set of observations concern what may prove to be the most salient response mechanism of them all: the content screening-and-removal policies of the platforms themselves, as expressed and established via their terms-of-service (TOS) agreements.

TOS agreements are the single most important documents governing digital speech in today's world, in contrast to prior ages where the First Amendment provided the road map for speech that was permissible in public discourse.[288] Today's most important speech fora, for better or worse, are online platforms, not public fora like public parks or streets. TOS agreements of private companies determine if speech on the major platforms is visible, prominent, or viewed, or if instead it is hidden, muted, or never available at all.[289] TOS agreements thus will be primary battlegrounds in the fight to minimize the harms that deep fakes may cause. The First Amendment has little to say about the choices that private companies make about what speech can and cannot appear on their services.

Some TOS agreements already ban certain categories of content. For instance, Twitter has long banned impersonation, without regard to the

---

288.   *See* Citron & Richards, *supra* note 45, at 1362.
289.   *See* Klonick, *supra* note 192, at 1630–38.

technology involved in making the impersonation persuasive.290 And Google's policy against non-consensual pornography now clearly applies to deep fakes of that kind. These are salutary developments, and other platforms can and should follow their lead even as all the platforms explore the question of what other variants of deep fakes likewise should be the subject of TOS prohibition.

As the platforms explore this question, though, they should explicitly commit themselves to what one of us (Citron) has called "technological due process."291 Technological due process requires companies be transparent—not just notionally but in real practical terms—about their speech policies. Platforms should be clear, for example, about what precisely they mean when they ban impersonation generally and deep fakes specifically. In our view, platforms should recognize that some deep fakes are not on balance problematic and should remain online. Thus, TOS should specify that deep-fake ban would not cover satire, parody, art, or education, as explored above. In our view, such deep fakes should not normally be filtered, blocked, muted, or relegated to obscurity.

Platforms should provide accountability for their speech-suppression decisions, moreover. Users should be notified that their (alleged) deep-fake posts have been blocked, removed, or muted and given a meaningful chance to challenge the decision.292 After all, as we noted above there is a significant risk that growing awareness of the deep fake threat will carry with it bad faith exploitation of that awareness on the part of those who seek to avoid accountability for their real words and actions via a well-timed allegation of fakery.

The subject of technological due process also draws attention to the challenge of just how platforms can and should identify and respond to content that may be fake. For now, platforms must rely on users and in-house content moderators to identify deep fakes. The choice between human decision-making and automation is crucial to technological due process.293 Exclusive reliance on automated filtering is not the answer, at least for now, because it is too likely to be plagued both by false positives and false negatives.294 It may have a useful

---

290.   CITRON & JURECIC, PLATFORM JUSTICE, *supra* note 46 at 14; *see also* CITRON, HATE CRIMES IN CYBERSPACE, *supra* note 40, at 228–42 (calling for platforms to adopt speech rules and procedures that provide greater transparency and accountability).

291.   Citron, *Technological Due Process*, *supra* note 245. Kate Klonick takes up this model in her groundbreaking work on the speech rules and practices of content platforms who she calls the "New Speech Governors." Klonick, *supra* note 192, at 1668–69.

292.   Note that 17 U.S.C. § 512(g) (2012) (part of the Digital Millennium Copyright Act) includes a provision requiring notice where an entity removes content based on a copyright infringement concern. Our proposal is not limited to copyright-infringement takedowns and would apply to muting or other forms of suppression that reduce visibility without outright removal of the content. Crucially, we are also not suggesting that law require moderation practices that emulates technological due process. Instead, we invoke the concept as an analogy to commitments to transparency and accountability, one that would be adopted voluntarily in the market, not as a direct regulatory mandate.

293.   *See* CITRON & JURECIC, PLATFORM JUSTICE, *supra* note 46, at 17.

294.   *Cf.* Georgia Wells et al., *The Big Loophole that Left Facebook Vulnerable to Russia Propaganda*, WALL ST. J (Feb. 22, 2018), https://www.wsj.com/articles/the-big-loophole-that-left-

role to play in flagging specific content for further review by actual analysts, but normally should not serve as the last word or the basis for automatic speech-suppressive action (though an exception would be proper for situations in which content previously has been determined, with due care, to be fraudulent, and software detects that someone is attempting to post that identical content).

The good news—and we would like to end on such a note—is that some of the largest platforms do recognize the problem deep fakes present, and are beginning to take steps to respond. Facebook, for example, plans to emphasize video content to a growing degree and has stated that it will begin tracking fake videos.295 Also underway are efforts to emphasize videos from verified sources while also affirmatively deemphasizing ones that are not; this will not correspond perfectly with legitimate versus fake videos of course, but it might help to some degree, although at some cost to the ability of anonymous speakers to be heard via that platform.296 Much more will be needed, but the start is welcome.

CONCLUSION

Notwithstanding the adage about sticks-and-stones, words in the form of lies have always had the ability cause significant harm to individuals, organizations, and society at large. From that perspective, the rise of deep fakes might seem merely a technological twist to a long-standing social ill.

But another adage—that a picture is worth a thousand words—draws attention to what makes the deep-fake phenomenon more significant than that. Credible yet fraudulent audio and video will have a much-magnified impact, and today's social media-oriented information environment interacts with our cognitive biases in ways that exacerbate the effect still further. A host of costs and dangers will follow, and our legal and policy architectures are not optimally designed to respond. Our recommendations would help with that to some degree, but the problem to a considerable degree would still remain. A great deal of further creative thinking is needed. We hope to have spurred some it by sounding this alarm.

---

facebook-vulnerable-to-russian-propaganda-1519314265 [https://perma.cc/3HU9-HYV7] (reporting that YouTube mistakenly promoted a conspiratorial video falsely accusing a teenage witness to the Parkland school shooting of being an actor).

295. *Id.*
296*. Id.*