Boston University School of Law
Scholarly Commons at Boston University School of Law

Faculty Scholarship

7-2018

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

Danielle K. Citron
Boston University School of Law

Robert Chesney
University of Texas

Follow this and additional works at: https://scholarship.law.bu.edu/faculty_scholarship

 Part of the First Amendment Commons, Internet Law Commons, and the Privacy Law Commons

Recommended Citation
Danielle K. Citron & Robert Chesney, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, Draft (2018).
Available at: https://scholarship.law.bu.edu/faculty_scholarship/640

This Article is brought to you for free and open access by Scholarly Commons at Boston University School of Law. It has been accepted for inclusion in Faculty Scholarship by an authorized administrator of Scholarly Commons at Boston University School of Law. For more information, please contact lawlessa@bu.edu.

BOSTON UNIVERSITY

****DRAFT****

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 with it 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.

* James Baker Chair, University of Texas School of Law; co-founder of Lawfare.

** Morton & Sophia Macht Professor of Law, University of Maryland Francis King Carey School of Law; Affiliate Fellow, Yale Information Society Project; Affiliate Scholar, Stanford Center on Internet and Society. We thank Benjamin Wittes, Quinta Jurecic, Nathaniel Gleicher, Andreas Schou, Klon Kitchen, and Patrick Gray for helpful suggestions. We are grateful to Susan McCarty, Samuel Morse, Jessica Burgard, and Alex Holland for research assistance.

Electronic copy available at: https://ssrn.com/abstract=3213954

TABLE OF CONTENTS

Electronic copy available at: https://ssrn.com/abstract=3213954

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

BOBBY CHESNEY* AND DANIELLE CITRON**

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* piece incorporated a video entitled "This Is Why We March," including a visually arrested 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, with an altogether different effect. In the fake, the torn sheet is not a bullseye, but rather a copy of the Constitution of the United States. While at one level the 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.

It is no surprise that the image went viral. Soon a fake of the video appeared, though it was more obvious the images had been manipulated. Still, it 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 dog whistle reference for neo-Nazis thanks to the word's role in a Hitler Youth anthem).

---

* James Baker Chair, University of Texas School of Law; co-founder of Lawfare.

** Morton & Sophia Macht Professor of Law, University of Maryland Francis King Carey School of Law; Affiliate Fellow, Yale Information Society Project; Affiliate Scholar, Stanford Center on Internet and Society. We thank Benjamin Wittes, Quinta Jurecic, Nathaniel Gleicher, Andreas Schou, Klon Kitchen, and Patrick Gray for helpful suggestions. We are grateful to Susan McCarty, Samuel Morse, and Jessica Burgard for research assistance.

1. Alex Horton, *A Fake Photo of Emma González Went Viral on the Far Right, Where Parkland Teens Are Villains*, WASH. POST: THE INTERSECT (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.

2. Emma González, *Emma González on Why This Generation Needs Gun Control*, TEEN VOGUE: NEWS AND POLITICS (Mar. 23, 2018, 5:20 AM), https://www.teenvogue.com/story/emma-gonzalez-parkland-gun-control-cover?mbid=social_twitter.

Electronic copy available at: https://ssrn.com/abstract=3213954

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 quality of the fakes were poor (though audiences inclined to believe their message might disregard the red flags).

Even with those constraints, though, many believed the fake, and harm followed. Our national dialogue on gun control has suffered some degree of distortion; Emma has likely suffered some degree of anguish over the episode; her family and friends have surely incurred harassment beyond whatever might otherwise have been the case.

Falsified imagery, in short, has already exacted significant costs both at the individual and societal levels. But the situation is about to get worse. Much worse, as this Article shows.

Technologies for altering images, video, or audio (or even creating them from scratch) are maturing rapidly. As they ripen and diffuse, the problems illustrated by the Emma González episode will expand rapidly and generate significant policy and legal challenges. One need only imagine a fake video depicting an American soldier murdering an innocent civilian in an Afghan village or a candidate for office making an inflammatory statement the night before an election. Screenwriters are already building such prospects into their plotlines.[3] The real world will not lag far behind.

Predictably, pornographers have been early adopters of the relevant technology, interposing the faces of celebrities into sex videos. This has given rise to the label "deep fake" for such digitized impersonations. We use that label here more broadly, as a

3. *See, e.g.*, *Homeland: Like Bad at Things* (Showtime television broadcast Mar. 4, 2018); *Taken: Verum Nocet* (NBC television broadcast Mar. 30, 2018) (deep video fake); *The Good Fight: Day 464* (CBS television broadcast Apr. 29, 2018) (deep video fake of alleged Trump pee tape); *The Good Fight: Day 408* (CBS television broadcast Mar. 11, 2018) (deep audio fake).

Electronic copy available at: https://ssrn.com/abstract=3213954

shorthand for the full range of hyper-realistic digital falsification of images, video, and audio.

That 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 porn context, with great potential for harm.[4] We 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 that will make 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 upside of deep fakes include artistic exploration and educative value. The downsides of deep fakes 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, including the distortion of policy debates, the manipulation of elections, the erosion of trust in institutions, the exacerbation of social divisions, the damage to national security, and the disruption of international relations. And, in what might be understood as a "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, "active measures" in special contexts like armed conflict and covert action, and technology-driven market responses, including not just promotion of debunking technologies, but also the prospect of privacy-destructive life logging as an alibi service. We end with a summary of our recommendations and warnings.

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.[5] The end result is realistic-looking

4. *See, e.g.*, Samantha Cole, *We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now*, VICE: MOTHERBOARD (Jan. 28, 2018, 1:13 PM), https://motherboard.vice.com/en_us/article/bjye8a/reddit-fake-porn-app-daisy-ridley. *See also* BuzzFeed (@BuzzFeed), *You Won't Believe What Obama Says In This Video!*, TWITTER (Apr. 17, 2018, 8:00 AM), https://twitter.com/BuzzFeed/status/986257991799222272.; *All Things Considered: Technologies to Create Fake Audio and Video Are Quickly Evolving*, (Tim Mak, Nat'l Pub. Radio broadcast Apr. 2, 2018), https://www.npr.org/2018/04/02/598916380/technologies-to-create-fake-audio-and-video-are-quickly-evolving; Julian Sanchez (@normative), TWITTER (Jan. 24, 2018, 1:26 PM) ("The prospect of any Internet rando being able to swap anyone's face into 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.").

5. *See* Cole, *supra*.

Electronic copy available at: https://ssrn.com/abstract=3213954

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.

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.[6] The field of digital forensics has been grappling with the challenge of detecting digital alterations for some time.[7] Generally, forensic techniques are automated and thus less dependent on the human eye to spot discrepancies.[8] While the detection of doctored audio and video was once fairly straightforward,[9] an emergent wave 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.[10] Much as experience refines the brain's neural nodes, examples train the neural network system.[11] If the network processes a broad array of training examples, it should be able to create increasingly accurate models.[12] It is through

6. *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.

7. Doctored images have been prevalent since the advent of the photography. *See Photo Tampering Throughout History*, IZITRU.COM, http://pth.izitru.com (last visited May 7, 2018), The gallery is curated by FourandSix Technologies, Inc. *See about us*, FOURANDSIX.COM, http://www.fourandsix.com/about-us.

8. *See* Tiffany 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. *See also* Rick Gladstone, *Photos Trusted but Verified*, N.Y. TIMES: LENS (May 7, 2014), https://lens.blogs.nytimes.com/2014/05/07/photos-trusted-but-verified/ (describing the website IZITRU.COM, https://www.izitru.com, which is spearheaded by Dartmouth's Dr. Hany Farid. It allows users to upload photos to determine if they are fakes. The service is aimed at "legions of citizen journalists who want to dispel doubts that what they are posting is real").

9. *See* Steven Melendez, *Can New Forensic Tech Win War on AI-Generated Fake Images?*, FAST COMPANY: ROBOT REVOLUTION (Apr. 4, 2018), https://www.fastcompany.com/40551971/can-new-forensic-tech-win-war-on-ai-generated-fake-images.

10. Larry Hardesty, *Explained: Neural Networks,* MIT NEWS (Apr. 14, 2017), http://news.mit.edu/2017/explained-neural-networks-deep-learning-0414.

11. Natalie Walchover, *New Theory Cracks Open the Black Box of Deep Neural Networks*, WIRED: SCIENCE (Oct. 8, 2017, 7:00 AM), https://www.wired.com/story/new-theory-deep-learning/.

12. Will Knight, *Meet the Fake Celebrities Dreamed Up By AI*, MIT TECH. REV.: THE DOWNLOAD (Oct. 31, 2017, 5:20 PM), https://www.technologyreview.com/the-download/609290/meet-the-fake-celebrities-dreamed-up-by-ai/.

Electronic copy available at: https://ssrn.com/abstract=3213954

this process that neural networks categorize audio, video, or images and generate realistic impersonations or alterations.[13]

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 said at the time.[14] 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 actually said.[15]

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.[16] One network, known as the generator, draws on a dataset to produce a sample that mimics the dataset.[17] The other network, the discriminator, assesses the degree to which the generator succeeded.[18] In an iterative fashion, the assessments of the discriminator inform the assessments of the generator. The result far exceeds the speed, scale, and nuance of what human reviewers could achieve.[19] Growing sophistication of the GAN approach is sure to lead to the production of increasingly convincing and nearly impossible to debunk deep fakes.[20]

---

13. Will Knight, *Real or Fake? AI is Making it Very Hard to Know,* MIT TECH. REV. (May 01, 2017), https://www.technologyreview.com/s/604270/real-or-fake-ai-is-making-it-very-hard-to-know/.

14. James Vincent, *New AI Research Makes It Easier to Create Fake Footage of Someone Speaking*, THE VERGE (July 12, 2017, 2:21 PM), https://www.theverge.com/2017/7/12/15957844/ai-fake-video-audio-speech-obama, Supasorn Suwjanakorn 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).

15. Charles Q. Choi, *AI Creates Fake Obama,* IEEE SPECTRUM (July 12, 2017, 2:00 PM), https://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/ai-creates-fake-obama; *see also* Joon Son Chung et al., *You Said That?*, British Machine Vision Conf. 2017 Conf. Paper (Feb. 2017), https://arxiv.org/abs/1705.02966.

16. *See* Ian Goodfellow et al., *Generative Adversarial Nets*, Int'l Conf. on Neural Info. Processing Sys. Conf. Paper (June 2014), https://arxiv.org/abs/1406.2661 (introducing the GAN approach). *See also* Tero Karras, et al., *Progressive Growing of GANs for Improved Quality, Stability, and Variation*, ICLR 2018 Conf. Paper (Apr. 2018), available at http://research.nvidia.com/sites/default/files/pubs/2017-10_Progressive-Growing-of/karras2018iclr-paper.pdf.

17. Karras, *supra* note 16.

18. *Id.*

19. *Id.*

20. Consider research conducted at Nvidia. Karras, *supra* note 16, at 2 (explaining novel approach that begins training cycle with low-resolution images and gradually shifts to higher-resolution images, producing better and much quicker results). The *New York Times* recently profiled Nvidia team's work. *See, e.g.*, 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. For further illustrations of the GAN approach, see Chris Donohue, *Implementation of Semantically Decomposing the Latent Spaces of Generative Adversarial Networks,* https://github.com/chrisdonahue/sdgan; Alec Radford et al., *Unsupervised Representation Learning with*

Electronic copy available at: https://ssrn.com/abstract=3213954

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,[21] Baidu's DeepVoice,[22] and GAN models.[23] Startup Lyrebird has posted short audio clips simulating Barack Obama, Donald Trump, and Hillary Clinton discussing its technology with admiration.[24]

This brief description reflects what is publicly known about private and academic efforts to develop deep-fake technology. Governmental research is currently a known unknown.[25] Classified research might lag behind commercial and academic efforts, but the reverse may be true. Given the possible utility of deep-fake techniques for various government purposes—not to mention the corresponding need to defend against hostile

---

Deep Convolutional Generative Adversarial Networks (2015), https://arxiv.org/abs/1511.06434; Martin Arjovsky et al., *Wasserstein GAN* (2017), https://arxiv.org/pdf/1701.07875.pdf**;** Phillip Isola et al., *Image-to-Image Translation with Conditional Adversarial Nets* (2016), https://phillipi.github.io/pix2pix/; Jun-Yan Zhu et al., *Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks*, https://junyanz.github.io/CycleGAN/**.**

21. Aaron van den Oord et al., *WaveNet: A Generative Model for Raw Audio,* ARXIV (Sept. 19, 2016), https://arxiv.org/pdf/1609.03499.pdf.

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

23. *See, e.g.,* Chris Donahue et al., *Synthesizing Audio with Generative Adversarial Networks*, ARXIV (Feb. 12, 2018), https://arxiv.org/pdf/1802.04208.pdf; Yang Gao et al., *Voice Impersonation Using Generative Adversarial Networks*, ARXIV (Feb. 19, 2018), https://arxiv.org/abs/1802.06840.

24. 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/. See Lyrebird's demo of generated voices of public figures reading tweets from their Twitter feeds at https://lyrebird.ai/demo/. The ability to cause havoc—via Twitter or otherwise—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 an inevitability and society benefits from its gradual introduction to it. https://lyrebird.ai/ethics/ ("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.").

25. 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." David Gunning, *Media Forensics (MediFor)*, DEFENSE ADVANCED RESEARCH PROJECTS AGENCY, https://www.darpa.mil/program/media-forensics (last visited May 7, 2018). 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 RESEARCH PROJECTS ACTIVITY, https://www.iarpa.gov/index.php/research-programs/diva (last visited May 7, 2018). 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, to Samuel Morse (Apr. 6, 2018, 7:49 AM) (on file with authors).

Electronic copy available at: https://ssrn.com/abstract=3213954

uses—it is a safe bet that state actors are conducting classified research in this area. 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.[26] Deep fakes have characteristics that ensure their spread beyond corporate or academic circles.[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: plutonium or highly enriched uranium to create nuclear weapons demonstrate the point.[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 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 or, worse, may have commercial or intellectual motivation to disperse it.[33]

---

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

27. *See* Dunaway, *supra.*

28. BENJAMIN WITTES & GABRIELLA BLUM, THE FUTURE OF VIOLENCE: ROBOTS, 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.

30. *See generally* GRAHAM 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 propagate the WannaCry cyber attacks, which wreaked havoc on network systems around the globe, threatening to erase files if a ransom wasn't paid through Bitcoin.

Electronic copy available at: https://ssrn.com/abstract=3213954

Bearing this in mind, the capacity to generate deep fakes is sure to diffuse rapidly. The capacity to generate effective deep fakes does not depend on scarce tangible inputs, but rather on access to knowledge like GANs and other approaches to machine learning. Governments may be engaged in classified research in this area, but the volume and sophistication of publicly available academic research and commercial services will ensure the steady diffusion of deep-fake capacity no matter efforts to safeguard it.[34]

The diffusion will reach beyond experts as user-friendly tools are developed and propagated online both in traditional and non-traditional commercial venues. This occurred with the advent graphic manipulation tools like Photoshop and malware like DDoS tools.[35] User-friendly capacity to generate deep fakes will emerge, especially with the ever-expanding market for cybercrime as a service on the dark web.[36]

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.[37] Now available for download is Fake App, "a desktop tool for creating realistic face swapping videos with machine learning."[38] 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 1,841 videos of himself.[39] 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.[40] 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, audience. As this section explores, networked phenomena, rooted in cognitive bias, will fuel that effort.[41]

---

34. *Supra* note at 33.

35. ARMOR, THE BLACK MARKET REPORT: A LOOK INSIDE THE DARK WEB 2 (Mar. 2018), https://www.armor.com/app/uploads/2018/03/2018-Q1-Reports-BlackMarket-DIGITAL.pdf (explaining that DDoS attack against organization can be purchased for $10/hour, or $200/day).

36. *Id.*

37. Emma Grey Ellis, *People Can Put Your Face on Porn—And the Law Can't Help You*, WIRED (Jan. 26, 2018, 7:00 AM), https://www.wired.com/story/face-swap-porn-legal-limbo.

38. FAKEAPP, https://www.fakeapp.org (last visited May 8, 2018).

39. 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.

40. *Id.*

41. *See generally* DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE (2014) (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 (2011); CASS SUNSTEIN, REPUBLIC.COM (2001).

Electronic copy available at: https://ssrn.com/abstract=3213954

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.[42] 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.[43]

The information revolution has disrupted this content distribution model.[44] 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.[45] This reduces the overall amount of gatekeeping, though control still remains with the companies responsible for our digital infrastructure.[46] For instance, content platforms have terms-of-service agreements, which ban certain forms of content based on companies' values.[47] They experience pressure from, or adhere to legal mandates of, governments to block or filter certain information like hate speech or "fake news."[48]

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, suppression of facts or opinions that some authority deems undesirable, or quality.

Content not only can find its way to online audiences, but can circulate far and wide, sometimes going viral. The interplay between cognitive heuristics (biases) and routine

---

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

43. Carr, *supra* note.

44. SIVA VAIDHYANATHAN, THE GOOGLIZATION OF EVERYTHING (AND WHY WE SHOULD WORRY) (2011).

45. 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. JACK GOLDSMITH & TIM WU, WHO OWNS THE INTERNET? (2008). The Internet visible in China is vastly different from the Internet visible in the EU, which is different the Internet visible in the United States.

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

47. CITRON, HATE CRIMES IN CYBERSPACE, *supra* note, at 232–35; Danielle Keats Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep*, 93 NOTRE DAME L. REV. 1035 (2018); Danielle Keats Citron & Benjamin Wittes, *The Internet Will Not Break: Fixing Section 230 Immunity for Bad Samaritans*, 86 FORDHAM L. REV. 401 (2017); Danielle Keats Citron & Helen Norton, *Intermediaries and Hate Speech: Fostering Digital Citizenship in the Twenty-First Century*, 91 B.U. L. REV. 1435 (2011); Danielle Keats Citron, *Cyber Civil Rights*, 89 B.U. L. REV. 61 (2009); *see also* Danielle Keats Citron & Quinta Jurecic, *Platform Justice* (on file with authors).

48. Citron, *Extremist Speech*, *supra* note, at. For important work on global censorship efforts, see the scholarship of Anupam Chander. *See, e.g.,* Anupam Chander, *Facebookistan*, 90 N.C. L. REV. 1807 (2012); Anupam Chander, *Googling Freedom*, 99 CALIF. L. REV. 1 (2011).

Electronic copy available at: https://ssrn.com/abstract=3213954

algorithmic practices make viral circulation possible. The following two phenomena--the "information cascade" dynamic and filter bubbles--make this possible.

First, consider the "information cascade" dynamic.[49] Information cascades are the result of the human tendency to credit what others know. Everyday interactions involve the sharing of information. Because people cannot know everything, they often rely on what others say even if it contradicts their own knowledge.[50] At a certain point, in other words, it is rational for people to stop paying attention to their own information and to look to what others know.[51] And when people pass along what others think, the credibility of the original claim snowballs.[52] Simply put, information cascades result when people stop paying sufficient attention to their own information and rely too much on what they assume others know.[53] They compound the problem by sharing the information onward, believing they have learned something valuable.[54] The cycle repeats, and the cascade strengthens.[55]

Social media platforms are a ripe environment for the formation of information cascades spreading content of all stripes and quality. From there, cascades can spill over to traditional mass-audience outlets, overcoming whatever gatekeeping that exists.[56] Social movements have leveraged the power of information cascades, from Black Lives Matter activists[57] to the Never Again movement of the Parkland High School students.[58] Arab Spring protesters spread videos and photographs of police torture, leading to the toppling of tyrants.[59] Journalist Howard Rheingold refers to positive information cascades as "smart mobs."[60] But not every mob is smart or laudable, and the information cascade dynamic does not account for such distinctions. The Russia covert action program to sow discord in the United States during the 2016 election provides ample demonstration.[61]

---

49. Carr, *supra* note. *See generally* DAVID EASLEY & JON KLEINBERG, NETWORKS, CROWDS, AND MARKETS: REASONING ABOUT A HIGHLY CONNECTED WORLD (2010); CASS SUNSTEIN, REPUBLIC.COM 2.0 (2007).

50. *See generally* Easley & Kleinberg, *supra* note, at.

51. *Id.*

52. *Id.*

53. *See generally* DAVID WEINBERGER, TOO BIG TO KNOW: RETHINKING KNOWLEDGE NOW THAT THE FACTS AREN'T THE FACTS, EXPERTS ARE EVERYWHERE, AND THE SMARTEST PERSON IN THE ROOM IS THE ROOM 84 (2011).

54. Citron, *supra* note, at 67.

55. *Id.*

56. *See generally* YOCHAI BENKLER, THE WEALTH OF NETWORKS: HOW SOCIAL PRODUCTION TRANSFORMS MARKETS AND FREEDOM (2006).

57. MONICA ANDERSON & PAUL HITLIN, PEW RESEARCH CTR., THE HASHTAG #BLACKLIVESMATTER EMERGES; SOCIAL ACTIVISM ON TWITTER (Aug. 15, 2016), http://www.pewinternet.org/2016/08/15/the-hashtag-blacklivesmatter-emerges-social-activism-on-twitter/.

58. 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.

59. Citron, *supra* note, at 68.

60. Rheingold, *supra* note.

61. The 2018 indictment of Internet Research Agency in the U.S. District Court for the District of Columbia is available at https://www.justice.gov/file/1035477/download; *see also* David Graham, *What the*

Electronic copy available at: https://ssrn.com/abstract=3213954

Compounding the problem is our natural tendency to propagate negative and novel information. Negative and novel information "grabs our attention as human beings and causes us to want to share that information with others—we're attentive to novel threats and especially attentive to negative threats."[62] Data scientists, for instance, studied news stories shared on Twitter from 2006 to 2010, using third-party fact-checking sites to generate a list of 126,000 rumors shared online.[63] According to the study, hoaxes and false rumors reached people ten times faster than accurate stories.[64] Even when researchers controlled for differences between accounts originating rumors, falsehoods were 70 percent more likely to get retweeted than accurate news.[65] The spread of fake news was not due to bots, which retweeted falsehoods at the same frequency as accurate information.[66] Rather, false news spread faster due to people retweeting inaccurate news items.[67] The study's authors hypothesized that falsehoods had greater traction because they seemed more "novel" and evoked more emotion than real news.[68] Fake tweets tended to elicit words associated with surprise and disgust, while accurate tweets included words associated with sadness and trust.

With human beings naturally primed to spread negative and novel falsehoods, automation can be weaponized to exacerbate those tendencies. Bots have been deployed to amplify and spread negative misinformation.[69] Facebook estimates that as many as 60 million bots may be infesting its platform.[70] Bots were responsible for a substantial portion of political content posted during the 2016 election.[71] Bots also can manipulate algorithms used to predict potential engagement with content.

Negative information not only is tempting to share, but it is also relatively "sticky." As social science research shows, people tend to credit—and remember—negative information far more than positive information.[72] Coupled with our natural predisposition towards certain stimuli like sex, gossip, and violence, that tendency

---

Mueller Indictment Reveals, THE ATLANTIC (Feb. 16, 2018), https://www.theatlantic.com/politics/archive/2018/02/mueller-roadmap/553604/; Tim Mak & Audrey McNamara, *The Russia Investigations: Mueller Indicts The 'Internet Research Agency*,' NAT'L PUB. RADIO (Feb. 17, 2018, 7:00 AM), https://www.npr.org/2018/02/17/586690342/mueller-indictment-of-russian-operatives-details-playbook-of-information-warfare.

62. *Id*.

63. Soroush Vosoughi et al., *The Spread of True and False News Online*, SCIENCE (Mar. 2018), http://science.sciencemag.org/content/359/6380/1146/tab-pdf.

64. *Id.*

65. *Id.*

66. *Id.*

67. *Id.*

68. *Id.*

69. Robinson Meyer, *The Grim Conclusions of the Largest Ever Study of Fake News*, THE ATLANTIC (Mar. 8, 2018) (quoting political scientist Dave Karpf).

70. Senate Judiciary Committee, Extremist Content and Russian Disinformation Online: Working with Tech to Find Solutions, Judiciary Committee, 2017.

71. David Lazer et al., *The Science of Fake News*, 359 SCIENCE 1094 (2018).

72. Elizabeth A. Kensinger, *Negative Emotion Enhances Memory Accuracy: Behavioral and Neuroimaging Evidence*, 16 CURRENT DIRECTIONS IN PSYCHOL. SCI. 213 (2007).

Electronic copy available at: https://ssrn.com/abstract=3213954

provides a welcome environment for harmful deep fakes.[73] The Internet amplifies this effect, which helps explain the popularity of gossip sites like TMZ.com.[74] Because search engines produce results based on our interests, they tend to feature more of the same—more sex and more gossip.[75] As social media researcher Danah Boyd explained:

> Our bodies are programmed to consume fat and sugar because they're rare in nature. . . . In the same way, we're biologically programmed to be attentive to things that stimulate: content that is gross, violent, or sexual, and gossip, which is humiliating, embarrassing, or offensive. If we're not careful, we're going to develop the psychological equivalent of obesity. We'll find ourselves consuming content that is least beneficial for ourselves and society as a whole.[76]

Second, filter bubbles further aggravate this state of affairs. 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.[77] Platforms' algorithms highlight popular information, especially if it has been shared by friends, surrounding us with content from relatively homogenous groups.[78] As endorsements and shares accumulate, the chances for an algorithmic boost increases. After seeing friends' recommendations online, individuals tend to share them with their networks.[79] Because people tend to share information with which they agree, social media users are surrounded by information confirming their preexisting beliefs.[80] This is what we mean by "filter bubble."[81]

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.[82]

---

73. PARISER, *supra* note, at 13–14.

74. Citron, *supra* note, at 68.

75. *Id.*

76. danah boyd, Web2.0 Expo, Streams of Content, Limited Attention: The Flow of Information Through Social Media (Nov. 17, 2009) (transcript available at http://www.danah.org/papers/talks/Web2Expo.html).

77. Citron, *supra* note, at 67.

78. *Id.*

79. *Id.* at 67.

80. *Id.* at 68.

81. 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 individuals visiting on average 33 articles from fake news websites), but narrow (group consuming fake news represented 10% of the public). *See* Andrew Guess et al., *Selective Exposure to Misinformation: Evidence From the Consumption of Fake News During the 2016 Presidential Campaign*, Jan 9, 2018 (published by Dartmouth College), https://www.dartmouth.edu/~nyhan/fake-news-2016.pdf.

82. Guess et al., *supra* note, at 11.

Electronic copy available at: https://ssrn.com/abstract=3213954

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. Nevertheless, deep fakes can inflict a remarkable array of harms, which are exacerbated by the 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 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.[83] 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.[84]

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

83. Emily Cruse, *Using Educational Video in the Classroom: Theory, Research, and Practice*, SAFARI MONTAGE (2013), http://www.safarimontage.com/pdfs/training/UsingEducationalVideoInTheClassroom.pdf.

84. 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*, 29th IEEE-CVPR 2016 Conf. Paper (June 2016), http://www.graphics.stanford.edu/~niessner/papers/2016/1facetoface/thies2016face.pdf.

Electronic copy available at: https://ssrn.com/abstract=3213954

hypothetical scenario under consideration. If done well, it would surely beat just having the professor asking students to imagine the shifting scenario out of whole cloth.[85]

The educational value of deep fakes will surely 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.[86] One can imagine deep fakes deployed to support educational campaigns like Mothers Against Drunk Driving, It Gets Better, or Eat Healthy.

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.[87]

2. Art

The potential artistic benefits of deep-fake technology certainly relate to its educational benefits, of course, but stand apart from them because they need not serve any formal educational purpose. Indeed, the benefits to creativity generally are already familiar to mass audiences thanks to the use of existing technologies to resurrect dead performers for fresh roles.[88] The startling appearance of what appeared to be 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. This was a prominent illustration that delighted some and upset others.[89] The *Star Wars* contribution to this theme continued in *The Last Jedi*, when the death of Carrie Fisher led the filmmakers to fake additional dialogue, using snippets from real recordings.[90]

Not all artistic uses of deep-fake technologies will have commercial potential. The possibilities are rich and varied. 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.

---

85. 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

86. *See supra* at note 8.

87. 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.

88. 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.

89. 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.

90. Evan Narcisse, *It Took Some Movie Magic to Finish 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.

Electronic copy available at: https://ssrn.com/abstract=3213954

3. *Autonomy*

Just as art overlaps with education, deep fakes implicate self-expression. Not all uses of deep fake capacity for self-expression are best understood as art. For example, deep-fake technology may be used to facilitate "avatar" experiences for individuals towards a variety of self-expressive ends that might best be described in terms of autonomy.

A striking example was an article suggesting that individuals suffering from certain physical disabilities could 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.[91] The utility of deep-fake technology for any such avatar experience, which need not be limited to sex, closely relates to more familiar examples of technology like video games that enable a person to have or perceive experiences that might otherwise be impossible, dangerous, or otherwise undesirable if pursued in person. 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. The customizable avatars from Nintendo Wii (known as a "Mii," naturally) provide a familiar and non-threatening (if simple) example.

Deep-fake technology will confer 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, the capacity to produce 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 in which they hide the liar's creative role. Deep fakes will emerge as powerful mechanisms for some to exploit and sabotage others.

91. 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.

Electronic copy available at: https://ssrn.com/abstract=3213954

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 those 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 in any event that any debunking would fail to reach far and fast enough to prevent or undo the initial damage.[92] 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.[93]

Not all value extraction takes a tangible form. Deep-fake technology can also be used to exploit peoples' sexual identities for others' gratification.[94] Thanks to deep-fake technology, peoples' faces, voices, and bodies can be swapped into real pornography.[95] A subreddit (now closed) featured deep-fake sex videos of female celebrities, amassing more than 100,000 users.[96] As one Reddit user asked, "I want to make a porn video with my ex-girlfriend. But I don't have any high-quality video with her, but I have lots of good photos."[97] 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.[98]

These examples highlight an important point: the gendered dimension of deep-fake exploitation. In all likelihood, the majority of victims of fake sex videos will be female.

92. ADAM DODGE & ERICA JOHNSTONE, DOMESTIC VIOLENCE ADVISORY: USING FAKE VIDEO TECHNOLOGY TO PERPETUATE INTIMATE PARTNER ABUSE 6 (Apr. 25 2018), http://withoutmyconsent.org/blog/new-advisory-helps-domestic-violence-survivors-prevent-and-stop-deepfake-abuse. The advisory was published by the non-profit organization Without My Consent, which combats online invasions of privacy.

93. 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., BROOKINGS INST., SEXTORTION: CYBERSECURITY, TEENAGERS, AND REMOTE SEXUAL ASSAULT (May 11, 2016), https://www.brookings.edu/wp-content/uploads/2016/05/sextortion1-1.pdf.

94. 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/; DODGE & JOHNSTONE, *supra* note, at 6 (explaining the likelihood that domestic abusers and cyber stalkers will use deep sex tapes to harm victims).

95. Danielle Keats Citron, *Cyber Sexual Exploitation*, U.C. IRVINE L. REV. (forthcoming).

96. Dodge & Johnstone, *supra* note, at 7.

97. Dodge & Johnstone, *supra* note, at 7.

98. *Id.*

Electronic copy available at: https://ssrn.com/abstract=3213954

This has been the case for cyber stalking and non-consensual pornography, and likely will be the case for deep sex fakes.[99]

One can easily imagine deep-fake videos subjecting individuals to violent, humiliating sex acts. This shows that not all such fakes will be designed primarily, or at all, for the sexual or financial gratification of the creator. Some will be nothing less than cruel weapons meant to inflict pain.

When victims discover that they have been used in fake sex videos, the psychological damage may be profound—whether or not this was the aim of the creator of the video. Victims may feel humiliated and scared. Fake sex videos force individuals into virtual sex, reducing them to sex objects. As Robin West has astutely observed, threats of sexual violence "literally, albeit not physically, penetrate women's bodies."[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]

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 various 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.

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 even embarrassing behavior like sounding incoherent. Depending on the circumstances, timing, and circulation of the fake, the effects could be devastating. 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.

---

99. ASIA EATON ET AL., CYBER CIVIL RIGHTS INITIATIVE, 2017 NATIONWIDE ONLINE STUDY OF NON-CONSENSUAL PORN VICTIMIZATION AND PERPETRATION 12 (June 2017), https://www.cybercivilrights.org/wp-content/uploads/2017/06/CCRI-2017-Research-Report.pdf ("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]).

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

101. Deep sex fakes should be understood as part of the broader cyber stalking phenomenon, which more often targets women and usually involves online assaults that are sexually threatening and sexually demeaning. *See* CITRON, HATE CRIMES, *supra* note, at.

Electronic copy available at: https://ssrn.com/abstract=3213954

Even if the victim has the ability to debunk the fake via alibi or otherwise, that fix 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 involves millions of dollars) once cautious teams make others picks and the victims falls into the later rounds of the draft (or out of the draft altogether).[102]

The nature of today's communications environment enhances the capacity of deep fakes to cause reputational harm. The combination of cognitive biases and algorithmic boosting described above 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. Ever-improving search capacities combine with these considerations to increase the chances that potential employers, business partners, or romantic interests will encounter the fake.

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]

---

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/.

103. CAREERBUILDER: PRESS ROOM, NUMBER OF EMPLOYERS USING SOCIAL MEDIA TO SCREEN CANDIDATES AT ALL-TIME HIGH, FINDS LATEST CAREERBUILDER STUDY (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.

104. This has been the case for nude photos posted without consent, often known as revenge porn. CITRON, HATE CRIMES IN CYBERSPACE, *supra* note, at 17–18, 48–49.

105. 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. Nicola Fox Hamilton, *Romantic Relationships and Online Dating, in* APPLIED CYBERPSYCHOLOGY 144–60 (Alison Attrill & Chris Fullwood, eds., 2016), https://link.springer.com/chapter/10.1057/9781137517036_9; Madeleine Holden, *How to Use Bumble to Guarantee Yourself a Date*, ASKMEN, https://www.askmen.com/dating/dating_advice/how-to-use-bumble-to-guarantee-you-a-date.html (last visited May 10, 2018).

In 2012, Match, E-harmony, JDate, and Christian Mingle announced that they would begin conducting background checks on all potential users for crimes of violence, sexual assault, and identity theft. This

Electronic copy available at: https://ssrn.com/abstract=3213954

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.

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 possibilities:

- 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 horrific things that they did not.[107]
- 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.
- 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.

---

came after Match was sued in 2011 for negligence after one user was raped while on a date with a repeat sex offender she met through the site.

Tim Newcomb, *Major Online Dating Sites to Start Background Checks on Users,* TIME (Mar. 21, 2012), http://newsfeed.time.com/2012/03/21/major-online-dating-sites-to-start-background-checks-on-users/. In 2017, the online dating site 'Gatsby' began blocking all possible users that have been convicted of any crime. Gatsby's founder Joseph Penora said that the checks take only 3.2 seconds to conduct. They update their data by conducting monthly scans of all their users. At this time, industry giants like Tinder, Grindr, and Bumble offer no such protection for their users. Rachel Thompson, *New Dating App Goes Full Blown "Law and Order," Bans Everyone with a Criminal Record,* MASHABLE (May 23, 2017), https://mashable.com/2017/05/23/gatsby-dating-app/?utm_cid=hp-n-1#yDeOtq4zHEqy.

107. Linton Weeks, *A Very Weird Photo of Ulysses S. Grant,* NAT'L PUB. RADIO (Oct. 27, 2015), https://www.npr.org/sections/npr-history-dept/2015/10/27/452089384/a-very-weird-photo-of-ulysses-s-grant.

Electronic copy available at: https://ssrn.com/abstract=3213954

- False audio might convincingly depict U.S. officials privately "admitting" a plan to commit an outrage overseas, exquisitely 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.

As these hypotheticals suggest, the threat posed by deep fakes has 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. Sometimes the lies are intended to undermine the credibility of participants in such debates, and sometimes the lies erode the factual foundation that ought to inform policy discourse. Even without prevalent deep fakes, information pathologies are abound. But deep fakes will make the situation worse by raising the stakes for the "fake news" phenomenon in dramatic fashion (quite literally).[108]

Many actors will have sufficient interest to exploit the capacity of deep fakes to skew information and thus manipulate beliefs. Some people will do it for reasons of state, as in the case of the Russian Government.[109] Other folks will do it as a form of unfair intellectual competition in the battle of ideas. And other individuals will do it simply as a tactic of intellectual vandalism. 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 explore in the next section.

One of the prerequisites for democratic discourse is a shared universe of facts and truths supported by empirical evidence.[110] In the absence of an agreed upon reality, efforts to solve national and global problems will become enmeshed in needless first-order questions like whether climate change is real.[111] The large scale erosion of public faith in data and statistics has led us to a point where the simple introduction of empirical evidence can alienate those who have come to view statistics as elitist.[112] Effective deep

108. *Cf.* Franklin Foer, *The Era of Fake Video Begins*, THE ATLANTIC (May 2018), https://www.theatlantic.com/magazine/archive/2018/05/realitys-end/556877/.

109. 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.

110. Mark Verstraete & Derek E. Bambauer, *Ecosystem of Distrust*, 16 FIRST AMEND. L. REV. 129, 152 (2017).

111. *Id.* at 144.

112. *Id.*

Electronic copy available at: https://ssrn.com/abstract=3213954

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 cannot proceed on a sustained basis.

b. Manipulation of Elections

In addition to the ability of deep fakes to inject mistaken beliefs about questions of policy in the electoral process, deep fakes can enable a particularly disturbing form of sabotage: distribution of a damaging, but false, video or audio about a candidate. The potential to sway the outcome of an election is quite 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 example of the 2017 election in France illustrate the perils. In a variant of the operation executed against the Clinton campaign in the United States in 2016, the Russians mounted a covert-action program blending cyber-espionage and information manipulation in an effort to prevent the election of Emmanuel Macron as President of France in 2017. The campaign included the theft of large numbers of digital communications and documents, the alteration of some of them in hopes of making them seem problematic in various ways, and the dumping of the lot on the public accompanied by aggressive spin. The effort ultimately fizzled for many reasons, including: poor tradecraft making 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, leaving enough time for the Macron team and the media to discover and point out all these flaws.[113]

It was a bullet dodged, yes, but a bullet nonetheless. The Russians might have acted with greater care, both in terms of timing and tradecraft. They might have produced a more-damning fake document, for example, dropping it just as polls opened. Worse, they might have distributed a deep fake consisting ofseemingly-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, or at least it will not stay so limited. 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

113. *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.

Electronic copy available at: https://ssrn.com/abstract=3213954

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 whom 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. Religious institutions are an obvious target, as are entities ranging from Planned Parenthood to the NRA.[114] 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.

d. Exacerbating Social Divisions

The institutional examples relate closely to significant cleavages involving identity and policy commitments in American society. Indeed, this is what makes institutions attractive targets for falsehoods. As divisions grow and 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—no doubt grows.[115] However, institutions will not be the only ones targeted with deep fakes. We anticipate that deep fakes will reinforce and exacerbate 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. That is, the Russians may have intended to enhance American social divisions as a general proposition, rendering us less capable of forming consensus on important policy

114. Recall the Project Veritas videos of Planned Parenthood officials edited to embarrass the organization. Imagine the potential for deep fakes along those lines.

115. B.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 (2015).

Electronic copy available at: https://ssrn.com/abstract=3213954

questions and thus more distracted by internal squabbles.[116] Texas is illustrative.[117] Russia promoted conspiracy theories about federal military power during the innocuous, "Jade Helm" training exercises.[118] Russian operators organized an event in Houston to protest radical Islam and a counter-protest of that event;[119] they promoted a Texas independence movement.[120] 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.[121] Consider 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, for example. Such an event would impose intangible costs by sharpening societal divisions, but also 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.[122] Now, false cries in the form of deep fakes go viral, fueled by the persuasive power of hyper-realistic evidence in conjunction with the

116. Dismiss, Distort, Distract, Dismay, https://www.ies.be/node/3689.

117. The CalExit campaign is another illustration of Russian disinformation campaign. *Russian Trolls Promoted California Independence*, BBC, http://www.bbc.com/news/blogs-trending-41853131.

118. Cassandra Pollock & Alex Samuels, *Hysteria Over Jade Helm Exercise in Texas Was Fueled by Russians, Former CIA Director Says*, TEX. TRIB. (May 3, 2018, 2:00 PM), https://www.texastribune.org/2018/05/03/hysteria-over-jade-helm-exercise-texas-was-fueled-russians-former-cia-/.

119. 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.

120. 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.

121. The pizzagate conspiracy theory is a perfect example. https://en.wikipedia.org/wiki/Pizzagate_conspiracy_theory. 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.

122. 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.").

Electronic copy available at: https://ssrn.com/abstract=3213954

distribution powers of social media.[123] The panic and damage Holmes imagined may be modest in comparison to the potential unrest and destruction created by a well-timed deep fake.

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. And they need not 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 warning to the public about an incoming ballistic missile.[124] Less widely noted, we saw purposeful attempts to induce panic when the Russian Internet Research Agency mounted a rather-sophisticated and well-resourced campaign to create the appearance of a chemical disaster in Louisiana and an Ebola outbreak in Atlanta.[125] There was real but limited harm in these cases. The stories spread only so far because they lacked evidence and because the facts were easy to check.

We will not always be so lucky as malicious attempts to spread panic grow. Deep fakes no doubt will be used towards that end and will provide an additional degree of credibility. Imagine if the Atlanta Ebola story had been backed by a compelling fake audio appearing to capture a phone conversation with the head of the CDC describing terrifying facts and calling for a cover-up to keep the public calm. Induced panic will be more damaging in the future.

f. Undermining Diplomacy

Deep fakes will also disrupt diplomatic relations and roil international affairs, especially where the fake is circulated publicly and has the effect of galvanizing public opinion. The recent Saudi-Qatar crisis might have been fueled by a hack in which someone injected fake stories with fake quotes by Qatar's emir into a Qatari news site.[126] The manipulator behind the lie can 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, of course. If entities are in a good position to make smart decisions about the weight to be given potential fakes, after all, it would be the intelligence agencies of the most-capable governments. 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

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

124. Cecilia Kang, *Hawaii Missile Alert Was'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.

125. Adrian Chen, *The Agency*, N.Y. TIMES (June 2, 2015), https://www.nytimes.com/2015/06/07/magazine/the-agency.html.

126. Krishnadev Calamur, *Did Russian Hackers Target Qatar?*, THE ATLANTIC (June 6, 2017), https://www.theatlantic.com/news/archive/2017/06/qatar-russian-hacker-fake-news/529359/.

Electronic copy available at: https://ssrn.com/abstract=3213954

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 options that the government may have 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 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. 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 considerable utility as a form of disinformation supporting strategic, operational, or even tactical deception. This is an ancient 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.[127] In that sense, deep fakes will be (or 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 continues to be the case in Afghanistan. In that context, we have long seen contending claims about civilian casualties—including, at times, the use of falsified evidence as well as allegations 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.

The realm of national security is broad by most measures, extending beyond military operations as such. But while definitional disputes abound—and are beyond the scope of this article—most would at least include the capabilities and activities of intelligence agencies in its scope. And here, too, deep fakes have the potential to cause considerable harm. 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 that this can

127. Jamie Rubin, *Deception: the Other 'D' in D-Day*, NBC NEWS: THE ABRAMS REPORT (June 5, 2004, 4:22 AM), http://www.nbcnews.com/id/5139053/ns/msnbc-the_abrams_report/t/deception-other-d-d-day/#.WvQt5NMvyT8.

Electronic copy available at: https://ssrn.com/abstract=3213954

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, or not far enough, in the case of the Snowden revelations is not our concern here. Our point is simply that this dynamic could be exploited if one wished to create distractions for an intelligence agency or even generate conditions that would lead a society to limit what that agency is authorized to do. None of that would be easily done, but with deep fakes the prospect of a strategic operation to bedevil a competing state's intelligence services becomes more plausible.[128]

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.[129] 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.[130]

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.[131]

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 in exactly this way. We already have seen many examples of "stings" pursued without the benefit of deep-fake technology.[132] Convincing deep fakes will make such stings more likely to succeed. Media

128. 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 then to embarrass NSA through a series of taunting public releases of those capabilities—with some degree of accompanying spin suggesting an interest in promoting doubt both in the U.S. and abroad about the wisdom of allowing 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.

129. 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).

130. *Id.* at 34.

131. 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.

132. *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-

Electronic copy available at: https://ssrn.com/abstract=3213954

entities may grow less willing to take risks in that 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. 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. First, 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, plainly, 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 a second and equally pernicious way. Ironically, this second approach will become more plausible 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 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 thereby 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,[133] the fact remains that it has made substantial inroads on 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

---

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.

133. 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.

Electronic copy available at: https://ssrn.com/abstract=3213954

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.[134]

The combination of *truth* decay and *trust* decay creates greater space for authoritarianism. Authoritarian regimes and leaders with authoritarian tendencies benefit when objective truths lose their power.[135] 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.[136]

Cognitive bias will reinforce these unhealthy dynamics. As Part II explored, people tend to believe facts that accord with our preexisting beliefs.[137] As research shows, people often ignore information that contradicts their beliefs and interpret ambiguous evidence as consistent with their beliefs.[138] People are also inclined to accept information that pleases them when given the choice.[139] Growing appreciation that deep fakes exist may provide a very convenient excuse for motivated reasoners to embrace these dynamics, even when confronted with information that is in fact true.

III. WHAT CAN BE DONE? A SURVEY OF TECHNICAL, LEGAL, AND MARKET RESPONSES

What can be done to ameliorate these harms? This Part reviews various possibilities. First, we explore the prospects for technological solutions that would facilitate the detection and debunking of deep fakes. Second, we describe current and potential criminal and civil liability. Third, we discuss the role of regulators. Fourth, we identify ways in which the government might respond to deep fakes with active measures. Last, 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.

---

134. 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 (2018), https://cms.edelman.com/sites/default/files/2018-01/2018%20Edelman%20Trust%20Barometer%20Global%20Report.pdf.

135. MILES BRUNDAGE, FUTURE OF HUMANITY INSTITUTE ET AL., THE MALICIOUS USE OF ARTIFICIAL INTELLIGENCE: FORECASTING, PREVENTION, AND MITIGATION REPORT 46 (Feb. 2018), https://www.eff.org/files/2018/02/20/malicious_ai_report_final.pdf.

136. *Id.*

137. Michela Del Vicario et al., *Modeling Confirmation Bias and Polarization*, 7 NATURE: SCIENTIFIC REPORTS no. 40,391 (2017), https://www.nature.com/articles/srep40391.

138. 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).

139. Shanto Iyengar et al., *Selective Exposure to Campaign Communication: The Role of Anticipated Agreement and Issue Public Membership*, 70 J. POL. 186 (2008).

Electronic copy available at: https://ssrn.com/abstract=3213954

A. Technological Solutions

Technology has given us deep fakes. Might it provide us with a robust capacity for debunking them when needed, thus limiting their harmful potential? An efficient and generally effective method for rapid detection of deep fakes would go far towards resolving this topic as a matter of pressing public-policy concern. Such software would have to keep pace with innovations in deep-fake technology to retain that efficacy, to be sure. But if such technology existed and could be deployed through social media platforms, the systemic harms described above would be reduced. This might not protect individuals from deep fakes involving narrow or even isolated distribution rather than distribution-at-scale through a social-media platform.[140] At least, the impact of harmful deep fakes might be cabined while beneficial uses could continue unabated.

Unfortunately, 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.[141] Those efforts, however, are tailored to particular products rather than video or audio technologies generally. Hence, their limited impact. Assuming these efforts can withstand efforts to undo them, they will have only limited use until one system (or a cluster of them) 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.

For now, we are left to seek a generally applicable technology that can detect manipulation in content without an expectation that the content comes with an internal certification. Dartmouth 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 deep fake ways to break the forensic system."[142] The defense, in short, is faring poorly at the moment 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 to create and to detect deep fakes. For example, growing awareness of the problem might produce the conditions needed for grantmaking agencies like the National Science Foundation (NSF) and the Defense Advanced Research Projects Agency (DARPA) to

140. Louise Matsakis, *Artificial Intelligence Is Now Fighting Fake Porn,* WIRED (Feb. 14, 2018), https://www.wired.com/story/gfycat-artificial-intelligence-deepfakes/. (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 "masks" the face of a possible deep fake, and searches the internet to see if the body and background footage exist elsewhere.).

141. For examples of provenance technologies in development, *see, e.g.*, Dia Kiyyali, Set Your Phone to ProofMode (Apr. 2017), https://blog.witness.org/2017/04/proofmode-helping-prove-human-rights-abuses-world/ (describing the concept of a metadata-rich "Proof Mode" app for Android devices).

142. *Id.*

Electronic copy available at: https://ssrn.com/abstract=3213954

begin steering funds towards scalable detection systems that can be commercialized or even provided freely. DARPA already 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.[143]

Those same conditions also might generate new market forces encouraging 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 significantly minimize the harms deep fakes might cause.

B. Legal Solutions

If technology alone will not save us, might 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 laws.

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.

The normativity of a flat ban is doubtful. Digital manipulation is not inherently problematic. Deep fakes exact significant harm in certain contexts but not all. A prohibition of deep fakes would prohibit routine modifications that improve the clarity of digital content. It would chill experimentation in a diverse array of 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. 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? Serious concerns remain even under that scenario.

The very existence of a general prohibition of deep fakes would cast a significant shadow, chilling expression crucial to self-governance and democratic culture. Government officials could use a deep-fakes ban to censor unpopular or dissenting views. 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

143. 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/ (""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.").

Electronic copy available at: https://ssrn.com/abstract=3213954

the ideological powers that be."[144] As James Weinstein wisely 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."[145]

Although self-serving prosecutions are not inevitable, they are a real risk. A dislike of minority or unpopular viewpoints, combined with ambiguity surrounding a deep-fake creator's intent, might chill public discourse.[146] The "'risk of censorious selectivity by prosecutors' will . . . distort perspectives made available" to the public.[147] 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, we are skeptical that a ban on deep fakes could withstand constitutional challenge. Deep fakes implicate freedom of expression, even though they involve intentionally false statements.[148] In the landmark 1964 decision *New York Times v. Sullivan*,[149] the Supreme Court held that false speech enjoys constitutional protection insofar as its prohibition would chill truthful speech.[150]

---

144. Thomas v. Collins, 323 U.S. 516, 545 (1945) (Jackson, J., concurring) ("Our forefathers did not trust government to separate the true from false for us."); Frank I. Michelman, *Conceptions of Democracy in American Constitutional Argument: The Case of Pornography Regulation*, 56 TENN. L. REV. 291, 302 (1989). Justice Oliver Wendell Holmes cautioned against the "human inclination to silence opinions that we dislike." Abrams v. United States, 250 U.S. 616, 624 (1919) (Holmes, J., dissenting). "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." 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.

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

146. *Id.* ("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 find that any misstatement of fact was an innocent one." *Id.*

147. *Id.* at 361.

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

149. N.Y. Times v. Sullivan, 376 U.S. 254 (1964) (imposing actual malice requirement on libel claims brought by public officials to ensure that public discussion could be robust, uninhibited, and wide open).

150. For a superb discussion of the constitutional significance of lies in the aftermath of *Alvarez*, see Alan Chen & Justin Marceau, *High Value Lies, Ugly Truths, and the First Amendment*, 68 VAND. L. REV. 1435, 1441 (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 two-level theory of the First Amendment that treats high value speech with stringent protections and second tier of speech that falls outside the First Amendment's coverage). The Court, in *Hustler Magazine v. Falwell*, struck down an intentional infliction of emotional distress claim based on a fake advertisement in the defendant magazine suggesting the Reverend Jerry Falwell lost his virginity to his mother. The Court refused to uphold the claim because there was no proof of actual malice—that defendant knew the advertisement was false or was reckless as to its truth or falsity. Hustler Magazine v. Falwell, 485 U.S. 46 (1988).

Electronic copy available at: https://ssrn.com/abstract=3213954

In 2012, in *United States v. Alvarez*,[151] the Court went even further. Eight Justices, writing in plurality and concurring opinions, concluded that "falsity alone" does not remove expression from First Amendment protection.[152] The Court held that barring the application of an exception to the First Amendment's protections, false statements can be regulated only insofar as defendants intend to cause "legally cognizable harm" and a direct causal link existed between the "restriction imposed and the injury to be prevented."[153]

This would seem to preclude a sweeping ban on deep fakes, yet it leaves considerable room for carefully tailored prohibitions of certain intentionally harmful deep fakes. As the Court acknowledged 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.[154] Some deep fakes fall into those categories. Categories of unprotected lies include the defamation of private persons, fraud, and impersonation of government officials.[155] For the same reason, speech integral to criminal conduct and the imminent-and-likely incitement of violence enjoy no First Amendment protection.[156]

Deep-fake bans particular to these scenarios have brighter prospects for surviving constitutional challenge. And, for much the same reasons, this rifle-shot approach to criminal and civil liability results in a more attractive balance of costs and benefits from the normative perspective. And so we turn now to a discussion of specific possibilities, starting with civil liability.

2. Specific Categories of Civil Liability

It may be that deep fakes cannot and should not be banned on a sweeping, generalized basis. But the question remains whether, in particular situations, their creators and distributors should be subject to civil liability for the harms they cause. In this section, we review the most relevant existing laws, as well as possibilities for new ones.

a. Threshold Obstacles

Before reviewing the prospects for particular theories of liability, we pause to emphasize a pair of devilish threshold problems.

---

151. United States v. Alvarez, 567 U.S. 709 (2012) (plurality opinion) (striking down the federal Stolen Valor Act, which punished people who claimed to have military honors that they did not; splintered decision on the matter of the level of scrutiny for government regulation of lies; unanimous on the notion that lies cannot be punished if no harm results).

152. *Alvarez*, 567 U.S. at 719.

153. *Id*. at 718, 743. The Justices were unanimous on the view that lies that cause no real harm are protected speech unless those lies concern narrow categories of speech that are not covered by the First Amendment. Chen & Marceau, *supra* note, at 1480.

154. Citron, *supra* note, at.

155. Chen & Marceau, *supra* note, at 1480.

156. Citron, *supra* note, at 165.

Electronic copy available at: https://ssrn.com/abstract=3213954

The first involves attribution. Civil liability cannot make a useful contribution to ameliorating the 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 provenance in connection with a deep fake might be insufficient to identify its creator. And it arises again at the next stage when the creator or someone else posts the deep fake on social media or otherwise injects it into the marketplace of information in a manner that others can access. A careful distributor of the deep fake may take pains to be anonymous, including but not limited to the use of technologies like Tor.[157] If so, the IP addresses connected to posts may be impossible to find and then trace back to the responsible parties.[158] 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 further circulation of the content.

A second obstacle arises when the creator of the deep fake—or the platform circulating it—are outside the United States and thus beyond the effective reach of U.S. legal process and, perhaps, in a jurisdiction where local legal action is unlikely to be effective. Even if attribution is not a problem (such as, where the creator is known or the plaintiff is pursuing a distributor), it still may be impossible to use civil remedies effectively. Of course, this general problem with civil liability could arise in any setting. But the global nature of online platforms makes it a particular problem in the deep-fake context.

Even if perpetrators can be identified and reside in the U.S., civil suits are expensive. Victims usually bear the costs of bringing civil claims, and those costs can be heavy.[159] They may be hesitant to spend scarce resources if posters are effectively judgment proof. Worse, a version of the "Streisand Effect" overhangs the decision to sue when, as is often the case, the deep fake is embarrassing or reputationally harmful. Lawsuits attract publicity, and unless the victim is permitted to sue under a pseudonym, filing the suit may exacerbate the victim's harm.[160]

All that said, these hurdles are not always fatal for would-be plaintiffs. If someone is able and willing to sue over a deep fake, then, the next question is whether current laws provide useful causes of action.

b. Suing the Creators of Deep Fakes

We start by considering the options for victims to sue the creator of a deep fake. There are several bodies of law that might come into play, including claims under the headings of both intellectual property and tort law.

157. Citron, *supra* note, at 165.

158. *Id.*

159. Citron, *supra* note, at 122 (exploring limits of civil law in redressing injuries resulting from cyber stalking).

160. 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, 1:36 AM), https://www.techdirt.com/articles/20050105/0132239_F.shtml.

Electronic copy available at: https://ssrn.com/abstract=3213954

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.[161] 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.[162]

The prospects for success, however, would be uncertain. The defendant will surely argue that the fake is a "fair use" of the copyrighted material, intended for educational, artistic, or other expressive purposes. Then too, whether the fake is sufficiently transformed from the original so as to earn "fair use" protection is a highly fact-specific inquiry, and we do not yet have a track record of courts grappling with such questions.[163]

Another prospect is the tort "right of publicity," which permits compensation for the misappropriation of someone's likeness for commercial gain.[164] For better or worse, 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.[165] 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-to-publicity claims concerning newsworthy matters on free-speech grounds.[166]

Tort law includes other concepts better suited to address some deep-fake scenarios. Most obviously, victims can sue for defamation where falsehoods are circulated either recklessly or negligently. Public officials and public figures, for their part, could sue posters for defamation if clear and convincing evidence exists of actual malice, that is, the defendant had knowledge the deep fakes were false or recklessly disregarded the possibility that they were false.[167] Private individuals only need to show that the defendant was negligent. Special damages do need not be shown if deep fakes injure someone's career, as would be the case if a video featured someone committing a crime or engaged in a sex video.[168] The closely related concept of suing for placing a person in a "false light"—that is, recklessly creating a harmful and false implication about someone in a public setting—likewise has clear potential for specific cases.[169]

---

161. Derek Bambauer, *Exposed*, 98 MINN. L. REV. 2025 (2014).

162. Megan Farokhmanesh, *Is It Legal to Swap Someone's Face Into Porn Without Consent?*, THE VERGE (Jan. 30, 2018, 2:39 PM), https://www.theverge.com/2018/1/30/16945494/deepfakes-porn-face-swap-legal (quoting Eric Goldman).

163. For an excellent discussion of this and other possibilities for suing deep-fake creators and purveyors, see Jesse Lempel, *Combatting Deep Fakes Through the Right of Publicity*, LAWFARE (Mar. 30, 2018, 8:00 AM), https://www.lawfareblog.com/combatting-deep-fakes-through-right-publicity.

164. JENNIFER ROTHMAN, RIGHT OF PUBLICITY: PRIVACY REIMAGINED FOR A PUBLIC WORLD (2018).

165. Lempel, *supra* note 164.

166. *Id.*

167. RESTATEMENT (SECOND) OF TORTS § 559 (AM. LAW INST. 1969).

168. *Id.*

169. *Id.*

Electronic copy available at: https://ssrn.com/abstract=3213954

Similarly, victims in some cases could sue for intentional infliction of emotional distress, which requires proof of "extreme and outrageous conduct."[170] Particularly humiliating content like deep-fake sex videos would amount to "extreme and outrageous conduct" because they would fall outside the norms of decency.[171]

Other privacy-focused torts seem relevant at first blush, yet are a poor fit on close inspection.[172] The "public disclosure of private fact" tort, for example, concerns the publication of private, "non-newsworthy" information that would highly offend the reasonable person.[173] Deep fakes certainly might cause such offense, but using a person's face in a deep-fake video does not amount to the disclosure of *private* information if the source image was generated from content posted online.[174] The intrusion-on-seclusion tort is likewise ill-suited to the deep-fake scenario. It narrowly applies to defendants who "intrude into a private place or invade a private seclusion that the plaintiff has thrown about his person or affairs."[175] Deep-fake videos do not involve invasions of spaces in which individuals have a reasonable expectation of privacy.

Current options for suing the creators of deep fakes are substantial in certain contexts, but limited in others. Subject to the practical obstacles noted above, civil liability is most robust in relation to defamation, false light, and intentional infliction of emotional distress, with some prospect for copyright infringement and right of publicity claims. In these respects, civil claims has potential to redress serious harm to specific individuals. But they cannot remedy a broad range of harms surveyed above.

Notably, civil claims, in particular, fall short with respect to certain systemic harms. Take the case of a creator of a deep fake that goes viral and, predictably, sets off violence in a community. The creator might be sued by a victim of that violence for negligence or perhaps even some form of intentional-tort liability, but the threshold hurdles described above loom large. Then again, it could be that for systemic harms, the important question is not whether it is possible to sue the creator of a particular deep fake, but rather whether it is possible to sue the platform that spread the fake widely and thus made possible its widespread ill effects.

c. Suing the Platforms

The preceding section suggests that individualized accountability for creators of harmful deep fakes will be possible in some cases but difficult in many others. The creators, however, are not the only parties that might bear responsibility. Given the key role that content platforms play in enabling the distribution of deep fakes, the most efficient way to mitigate the harm may be to impose liability on platforms. Content

170. Citron, *supra* note, at.

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

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

173. DANIEL J. SOLOVE & PAUL M. SCHWARTZ, PRIVACY LAW FUNDAMENTALS 42 (4th ed. 2017).

174. *Id.*

175. RESTATEMENT (SECOND) OF TORTS § 652B (AM. LAW INST. 1969).

Electronic copy available at: https://ssrn.com/abstract=3213954

platforms arguably are the cheapest cost avoiders because perpetrators may be difficult to find and deter.[176] In some contexts, content platforms may be the only realistic possibility for deterrence and redress.

Online platforms already have some incentive to play a screening role, thanks to the impact of moral suasion, market dynamics, and political pressures.[177] But as things currently stand, the prospect of civil liability does little to add to this pressure. This may seem odd at first glance. The preceding section explored several theories of liability available to victims of deep fakes, and these theories logically might be extended to the platforms in certain circumstances. All things being equal, that would be correct. All things are not equal, though.

In 1996, Congress provided platforms with a liability shield in the form of Section 230 of the Communications Decency Act (CDA). The basic idea is to make it hard for someone to sue an online platform based on its hosting of harmful content created or developed by others, with the exception of federal criminal law, the Electronic Communications Privacy Act, and intellectual property law.[178]

Section 230 accomplishes this in important ways. First, consider a situation in which an online platform displays content that is either republished from another source (such as quoting a news article) or generated by a user (such as a customer review posted on Yelp). Now imagine that this 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" of the problematic content. As courts have interpreted Section 230, online platforms enjoy immunity from liability for user-generated content even if they deliberately encourage the posting of that content.

Next, consider the situation where an online platform decides that is not merely going to enable users to post whatever they wish, but instead will engage in filtering or blocking to screen out 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. Simply put, it forbids civil suit against a platform based on good-faith act of filtering to screen out offensive content whether in the nature of obscenity, harassment, violence, or otherwise.[179]

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

176. Citron, *Mainstreaming Privacy Torts*, *supra* note.

177. Citron, *Extremist Speech, supra* note; Kate Klonick, *The New Governors: The People, Rules and Processes Governing Online Speech*, 131 HARV. L. REV. 1598 (2018). *See* CITRON, HATE CRIMES IN CYBERSPACE, *supra* note (exploring how and why content platforms moderate harmful content); Danielle Keats Citron & Helen Norton, *Intermediaries and Hate Speech: Fostering Digital Citizenship*, 91 B.U. L. REV. 1435 (2011).

178. Danielle Keats Citron & Benjamin Wittes, T*he Internet Will Not Break: Denying Bad Samaritans Section 230 Immunity*, 86 FORDHAM L. REV. 401 (2017). Lower courts have generally limited the intellectual property exemption to federal intellectual property claims, finding that state right to publicity claims enjoy immunity even though they are a variant of intellectual property law.

179. 47 U.S.C. § 230(e)(2).

Electronic copy available at: https://ssrn.com/abstract=3213954

"Internet."[180] Representative Chris Cox, for example, became interested after reading about a trial court decision holding Prodigy liable as a publisher of defamatory comments because it tried to filter profanity on its bulletin boards but did not incompletely.[181] A key goal 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.[182] At the time, online pornography was considered a scourge, and CDA sponsors like Senators James Exon and Slade Gorton were focused on making the "Internet" safe for kids.[183] Representatives Cox and Ron Wyden offered an amendment to the CDA—entitled "Protection for Private Blocking and Screening of Offensive Material"—that would become Section 230.[184] They argued that, "if this amazing new thing—the Internet—[was] going to blossom," companies should not be "punished for *trying* to keep things clean."[185]

This intent is clear in the language of Section 230(c)(2), which expressly concerns the scenario in which a platform might be deemed engaged in editorial activity based on filtering of offensive user-posted content. The other part of Section 230, however, lacks that narrowing language and as a result has proven to be a capable foundation for courts to interpret Section 230 immunity quite broadly.[186]

Section 230's immunity provision has been stretched considerably since its enactment, immunizing platforms even when they solicit or knowingly host illegal or tortious activity. The result has been a very permissive environment for hosting and distributing user-generated online content, yes, but also one in which it is exceptionally hard to hold providers accountable even in egregious circumstances—including situations in which someone is purveying systematic disinformation and falsehoods (state-sponsored or otherwise).[187]

---

180. JEFFREY KOSSEFF, THE TWENTY-SIX WORDS THAT GAVE US THE INTERNET (forthcoming Cornell Univ. Press).

181. 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), https://www.npr.org/sections/alltechconsidered/2018/03/21/591622450/section-230-a-key-legal-shield-for-facebook-google-is-about-to-change.

182. Danielle Keats Citron & Benjamin Wittes, *The Internet Will Not Break: Denying Bad Samaritans Section 230 Immunity*, 86 FORDHAM L. REV. 401 (2017).

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

184. H.R. REP. No. 104-223, Amendment No. 2-3 (1995) (proposed to be codified at 47 U.S.C. § 230).

185. Selyukh, *supra* note (quoting Cox).

186. Danielle Keats Citron, *Cyber Civil Rights*, 89 B.U. L. REV. 61, 118 (2009). In the landmark *ACLU v. Reno* decision, the Supreme Court struck down the CDA's blanket restrictions on Internet indecency under the First Amendment. Reno v. ACLU, 521 U.S. 844, 853 (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.

187. Tim Hwang, *Dealing with Disinformation: Evaluating the Case for CDA 230 Amendment*, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3089442.

Electronic copy available at: https://ssrn.com/abstract=3213954

Courts have taken this approach based on an assessment of "First Amendment values" that supposedly "drove the CDA."[188] Courts have extended the immunity provision so that it applies to a remarkable array of scenarios, including ones in which the provider republished content knowing it violated the law;[189] solicited illegal content while ensuring that those responsible could not be identified;[190] altered their user interface to ensure that criminals could not be caught;[191] and sold dangerous products.[192]

In this way, Section 230 has evolved into a kind of super-immunity that, among other things, prevents the civil liability system from incentivizing the best-positioned entities to take action against the most harmful content. This would have seemed absurd to the CDA's drafters.[193] 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.[194] Rebecca Tushnet put it well a decade ago: Section 230 ensures that platforms enjoy "power without responsibility."[195]

Unfortunately, that power now includes the ability to ignore the propagation of damaging deep fakes, even ones that cause specific and immediate harms such as sabotage to reputation, interference with business prospects, and the like. To be sure, there certainly are platforms that do not need civil liability exposure to take action against such obvious harms; market pressures and morals in some cases are enough. However, market pressures and morals are not always enough, and they should not have to be.

Should Section 230 be amended to make liability possible in a wider-range of circumstances? It has been done before in light of other harms. Recently so, in fact. In 2018, Congress enacted the Allow States and Victims to Fight Online Sex Trafficking Act (known as "FOSTA") to address the problem of websites facilitating various forms of sex

188. Jane Doe No. 1 v. Backpage.com LLC, 817 F.3d 12, 25 (1st Cir. 2016). The judiciary's insistence that the CDA reflected "Congress' desire to promote unfettered speech on the Internet" so ignores its text and history as to bring to mind Justice Scalia's admonition against selectively determining legislative intent in the manner of someone at a party who "look[s] over the heads of the crowd and pick[s] out [their] friends." ANTONIN SCALIA, A MATTER OF INTERPRETATION: FEDERAL COURTS AND THE LAW 36 (1997).

189. Shiamili v. Real Est. Group of N.Y., 952 N.E.2d 1011 (N.Y. 2011); Phan v. Pham, 182 Cal. App. 4th 323 (App. Ct. 2010).

190. 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. March 12, 2012).

191. Doe v. Backpage.com, LLC, 17 F.3d 12 (1st Cir. 2016).

192. *See, e.g.*, Hinton v. Amazon, 72 F. Supp. 3d 685, 687 (S.D. Miss. 2014).

193. 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. In his view, sites that solicit unlawful materials or have a connection to unlawful activity should not enjoy Section 230 immunity. *Id.*

194. Citron, *Cyber Civil Rights*, *supra* note, at 118; Mark Lemley, *Rationalizing Internet Safe Harbors* 6 J. TELECOMM. & HIGH TECH. L. 101 (2007); Douglas Gary Lichtman & Eric Posner, *Holding Internet Service Providers Accountable*, 15 SUP. CT. ECON. REV. 221 (2006).

195. Rebecca Tushnet, *Power Without Responsibility: Intermediaries and the First Amendment*, 77 GEO. WASH. L. REV. 101 (2009).

Electronic copy available at: https://ssrn.com/abstract=3213954

trafficking.[196] In relevant part, FOSTA added a new exception to the Section 230 immunity rule, comparable to the existing rule preserving the ability to sue for intellectual property claims. Now, plaintiffs, including state attorneys general, acting on behalf of victims, do not have to face 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, worrying about a slide towards greater liability exposure for online platforms and hence both decreased outlets and greater self-censorship among those remaining.[197] Others criticized FOSTA's indeterminate language because it could result in less filtering rather than more.[198] A crucial question is why Congress did not take the opportunity to reassess the immunity on a more wholesale basis, especially for platforms that can hardly be said to deserve it.

Which brings us back to the question whether Section 230 should be amended as to allow platforms to be held accountable for deep fakes. Building on a proposal that one of us (Citron) recently made with Benjamin Wittes, we argue that the answer is yes. In that prior work, Citron and Wittes argued for the benefits of Section 230(c)(1) to be made conditional, in contrast to being automatic under the status quo. Specifically, the entity would have to take "reasonable" steps to ensure that its platform is not being used for illegal ends:

> No provider or user of an interactive computer service that takes reasonable steps to address unlawful uses of its services shall be treated as the publisher or speaker of any information provided by another information content provider in any action arising out of the publication of content provided by that information content provider.[199]

To be sure, hard questions would arise regarding the metes and bounds of reasonableness in this setting. 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," yet "they may be able to deploy technologies to detect content previously deemed unlawful."[200] Inevitably, as Citron and Wittes observed, the "duty of care will evolve as technology improves."[201]

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 available technical and other means for detection and removal of harmful fakes are limited, so too will be the obligation

196. Danielle Citron & Quinta Jurecic, *FOSTA: The New Section 230 Amendment May Not End the Internet, But It's Not a Good Law Either*, LAWFARE (Mar. 28, 2018), https://www.lawfareblog.com/fosta-new-anti-sex-trafficking-legislation-may-not-end-internet-its-not-good-law-either.

197. *Id.* (discussing objections of Daphne Keller and Mike Godwin); see also Danielle Keats Citron & Quinta Jurecic, Platform Justice (forthcoming) (on file with authors).

198. Citron & Jurecic, *supra* note 200 (arguing that FOSTA is both too narrow and too broad).

199. Citron & Wittes, *supra* note, at.

200. *Id.*

201. *Id.*

Electronic copy available at: https://ssrn.com/abstract=3213954

on the part of the platform.[202] But as those means improve, so would the incentive to use them.[203]

We recognize that this proposal still runs risks, even if one accepts the bumps that come along with common law development of a novel standard of care. One might fear that opening the door to such liability will over-deter platforms due to a lack of certainty regarding the standard of care and to the prospect 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.[204] 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 akin to the excesses of 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, at which point the platform would have a specified window of time to examine and respond to the objection.

In the final analysis, a reasonably calibrated standard of care combined with such safeguards could reduce opportunities for abuses without interfering unduly with the further development of a vibrant internet or unintentionally turning innocent platforms into involuntary insurers for those injured through their sites. Approaching the problem as one of setting an appropriate standard of case more readily allows differentiating between different kinds of online actors, setting a different rule for websites designed to facilitate illegality from that applied to large ISPs linking millions to the Internet. That said, the cabining features that are needed to control the scope of platform liability ensure that this approach can be no more than a partial solution to the deep-fakes challenge. Other policy responses will be necessary.

---

202. 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. 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.

203. 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 threats, which requires an understanding of the context to determine its objectionable nature.

204. We see an example of that approach at several points in the history of the "Section 702" surveillance program.

Electronic copy available at: https://ssrn.com/abstract=3213954

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 gap identified above?

Only to a limited extent. While the criminal liability model in theory has the capacity to overcome some of the most significant limits on the civil liability model described above, its deterrent effect is different: being judgment proof might spare someone from fear of civil suit, but it is no protection from being sent to prison and bearing the other consequences of criminal conviction.[205] 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, this is a question of setting law enforcement priorities and allocating resources accordingly. Here, the track record is not promising. Notwithstanding notable exceptions, law enforcement, on the whole, has had a lackluster response to other forms of online abuse. In particular, state and local law enforcement often fail to pursue cyber stalking 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 cannot be scaled to anything like the same extent).[206] Although a wide range of deep fakes might warrant criminal charges, then, only the most extreme cases are likely to attract the attention of law enforcement.

Apart from questions of investigative and prosecutorial will, though, 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 are potentially relevant. If perpetrators post deep fakes in connection with the targeting of individuals, for example, they might violate the federal cyberstalking law, 18 U.S.C. 2261A, as well as analogous state statutes. Among other things, Section 2261A makes it a felony to use any "interactive computer service or electronic communication system" to "intimidate" a person in ways "reasonably expected to cause substantial emotional distress…"[207] This reflects the fact that, even when cyber stalking victims do not fear bodily harm, "their lives are totally

205. *Id*. at 123.

206. Citron, *supra* note, at 144. Assistant U.S. Attorney Mona Sedky is a shining example. *See The Lawfare Podcast: Mona Sedky and Benjamin Wittes on Prosecuting Sextortion*, LAWFARE INST. (June 25, 2016), https://www.lawfareblog.com/lawfare-podcast-mona-sedky-prosecutingsextortion [http://perma.cc/262G-KSLV].

207. 18 U.S.C. § 2261A(2) (2012). 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, at 124.

Electronic copy available at: https://ssrn.com/abstract=3213954

disrupted . . . in the most insidious and frightening ways."[208] 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.[209] 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, intimidate, threaten, or defraud" that person.[210] And while the "harm, intimidate, threaten" portion of such statutes to some extent tracks the cyberstalking statute described above, the 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 they were fake or if they were reckless as to their truth or falsity.[211] 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.

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.[212] 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

---

208. Reauthorization of the Violence Against Women Act: Hearing on S. 109-1033 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).

209. 18 U.S.C. § 2261A(2) (2012).

210. CAL. PENAL CODE § 528.5 (West 2009); N.Y. PENAL LAW § 190.25; MISS. CODE ANN. § 97-45-33; HAW. REV. STAT. ANN. § 711-1106.6; LA. REV. STAT. § 14:73.10; R.I. GEN. LAWS § 11-52-7.1; TEX. PENAL CODE § 33.07. The Texas impersonation statute withstood facial challenge in *Ex parte Bradshaw*, 501 S.W.3d 665 (Tex. App. 2016) (explaining that the conduct regulated by the statute is "the act of assuming another person's identity, without 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* 2017 Bill Text AZ H.B. 2489. 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 (2009) ("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 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, 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.

211. Eugene Volokh, *One to One Speech Versus One-to-Many Speech*, 107 NW. U. L. REV. 731 (2013).

212. YouTube, for example, barred incitement in 2008. *See* Peter Whorisky, *Youtube Bans Videos That Incite Violence*, WASH. POST (Sept. 12, 2008), http://www.washingtonpost.com/wp-dyn/content/article/2008/09/11/AR2008091103447.html.

Electronic copy available at: https://ssrn.com/abstract=3213954

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.[213] 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.[214] These experiments have run into constitutional hurdles, however, and for good reason.

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."[215] As the Court underscored in *Brown v. Hartlage*, the "State's fear that voters might make an ill-advised choice does not provide the State with a compelling justification for limiting speech."[216] Not surprisingly, courts therefore have struck down periodic attempts to ban election-related lies.[217] The entry of deep fakes into the mix will not likely change that result.

Criminal liability thus is not likely to be a particularly effective tool against deep fakes that pertain to elections. Setting aside the constitutional problems, moreover, 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

---

213. Multiple states prescribe criminal penalties for those who engage in similar conduct. *See, e.g.*, CAL. PENAL CODE § 404.6; VA. CODE ANN. § 18.2-408; FLA. STAT. ANN. § 870.01.; MONT. CODE ANN. § 45-8-105. 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).

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

215. Helen Norton, *Lies and the Constitution*, 2012 SUP. CT. REV. 161, 199.

216. 456 U.S. 45, 60 (1982). Although the Court has made some statements that suggest some room to sanction candidates' false speech, the Court has nonetheless struck down campaign regulation in each and every case in which the issue is raised. Stern, *supra* note, at 783. Free speech values support this result—given our distrust of government power to censor, political campaigns and elections are particularly fraught with potential for abuse. *See* Danielle Keats Citron & Neil Richards, *Four Principles for Digital Speech*, WASH. U. L. REV. (forthcoming). Couple the distrust of government censorship with the privileged treatment of political expression in the hierarchy of First Amendment freedoms and it is hard to imagine a criminal law sanctioning deep fakes in elections surviving strict scrutiny review.

217. *See, e.g.*, Susan B. Anthony List v. Dreihaus, 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) ("no 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").

Electronic copy available at: https://ssrn.com/abstract=3213954

not likely to end up in U.S. custody (thought it might mean something more to private actors through whom those agencies sometimes choose to act, at least if they intend to travel abroad).[218]

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.[219] 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.[220] For that matter, consider the full range of state actors (often a state's 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."[221] 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

218. 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 (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!).

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

220. 5 U.S.C. § 45(b).

221. 5 U.S.C. § 52.

Electronic copy available at: https://ssrn.com/abstract=3213954

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.[222] 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.[223] UDAP laws empower attorneys general to seek civil penalties, injunctive relief, and attorneys' fees and costs.[224]

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.[225] The FTC might contemplate asserting a role, under the rubric of "unfair and deceptive commercial practices," in response to the problem of fake news in general and deep-fake news in particular.[226] Any such efforts would face several obstacles, however. First, Section 230's immunity provision as currently written would protect platforms 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.[227] 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."[228] Finally, the prospect of a government entity attempting to distinguish real news from fake news—and suppressing

222. 15 U.S.C. § 45 (2012). 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). Under its unfairness authority, the FTC must show proof of harm, which is generally restricted to tangible injuries like economic and physical harm.

223. *See generally* Danielle Keats Citron, *The Privacy Policymaking of State Attorneys General*, 92 NOTRE DAME L. REV. 745, 755 (2016).

224. *See, e.g.*, California Unfair Business Act, CAL. BUS. & PROF. CODE § 17206 (West 2016) (imposing $2500 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").

225. 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/.

226. *See, e.g.,* 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.

227. *Id.*

228. *Id.*

Electronic copy available at: https://ssrn.com/abstract=3213954

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 21st century occupy a place in our information ecosystem similar to the central role that radio and television broadcasters enjoyed in the 20th century.[229] 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.[230] Amidst intense controversy, however, the FCC in late 2017 reversed course on this position on ISPs,[231] and in any event never asserted that so-called "edge providers" like Facebook also should be brought under the "telecommunication services" umbrella.[232]

As things currently stand, the FCC appears to lack jurisdiction (not to mention interest) over content circulated via social media. This could change, of course. 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 does regulate 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.

229. See Tim Wu, The Master Switch (2014).

230. Cite the 2015 Open Internet Order.

231. https://www.fcc.gov/document/fcc-releases-restoring-internet-freedom-order

232. https://www.fcc.gov/document/bureau-dismisses-petition-regulate-edge-provider-privacy-practices (2015).

Electronic copy available at: https://ssrn.com/abstract=3213954

Could that have at least some positive impact on deep fakes in the electoral setting? Perhaps. For outlets that come 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 the 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 citing the practical difficulties that would follow for small screens displaying even-smaller ads.

In the wake of the 2016 election, there is now pressure on the FEC to extend its reach to these platforms nonetheless, so that caveat might at some point drop out.[233] Even so, though, this certainly would not resolve the threat to elections posed by deep fakes.

There are two reasons for FEC regulation would not eliminate deep fakes' threat to elections. First, 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. Second, 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 government or non-state actor although not non-existent, as we have seen from time to time in the context of cybersecurity. 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, opposing politicians and

233. 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 spring 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.

Electronic copy available at: https://ssrn.com/abstract=3213954

electorates, local populations, allies, and so forth).[234] 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 U.S. or allied forces by depicting those forces burning a Koran 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 U.S. 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 State 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.[235] And it might entail the capture and detention of enemy personnel or supporters involved in such work.

---

234. The U.S. 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 the decision making of adversaries and potential adversaries while protecting our own." Joint Publication 3-13, Information Operations, at ix. Separately, it defines "military information support 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."Joint Publication 3-13.2, Military Information Support Operations, at vii. 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/.

235. 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 inflammatory deep fakes designed to cause an immediate violent response, though even there hard

Electronic copy available at: https://ssrn.com/abstract=3213954

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.[236] 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 U.S. 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. 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. Because covert action is not publicly acknowledged. Moreover, the diplomatic and political friction that might otherwise make a particular action unattractive is reduced in comparison to overt alternatives, athough 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 compliance with Constitution and statutes of the United States yet is conspicuously silent about international law, and many have speculated that this is construed within the government as domestic-law justification for activities that violate international law.[237]

---

questions would arise about the likely gap in time between creation of such a video and its actual deployment.

236. *See* 50 U.S.C. § 3093(e).

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

Electronic copy available at: https://ssrn.com/abstract=3213954

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 (so long as steps are taken to ensure that the targeted person or entity has at least some reason to believe that those costs are a response from the United States—a tricky but not insurmountable proposition where the sponsoring role of the United States is not meant to be acknowledged publicly).

Covert-action tools are not the only options the U.S. 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 U.S. 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 discretionwhen 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.[238]

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."[239] 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.

Is there currently a relevant national-emergency proclamation? There are a few possibilities. There are more than two-dozen currently active states of national emergency, as of the summer of 2018.[240] Most have little 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.

238. *See* 50 U.S.C. Chapter 35.

239. 50 U.S.C. § 1701(a).

240. 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), https://www.cnn.com/2017/08/12/politics/national-emergencies-trump-opioid/index.html. *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.

Electronic copy available at: https://ssrn.com/abstract=3213954

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."[241] 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…."[242] This was designed to allow for IEEPA sanctions against Russian entities that interfered in the 2016 election through means that included the DNC hack, and President Obama immediately used the authority to sanction Russia's FSB, GRU, and various other individuals and entities.[243] But could the same be done with respect to a foreign entity that engaged in no hacking, and instead focused entirely on using social media platforms to propagate false information in ways meant to impact American politics?[244]

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 under color of this framework.[245] IRA had engaged in extensive efforts to propagate false information into the American political debate, and when the Trump administration sanctioned it under color of the cyber executive order, this seemed an endorsement of the proposition that politically targeted information operations online were enough, even without hacking, to come within the scope of that IEEPA framework. A close read of the Treasury Department's explanation of IRA's inclusion, however, includes just enough reference to "misappropriation of

241. Exec. Order (Apr. 1, 2015), https://obamawhitehouse.archives.gov/the-press-office/2015/04/01/executive-order-blocking-property-certain-persons-engaging-significant-m.

242. Exec. Order No. 13757 (Dec. 28, 2016), https://www.treasury.gov/resource-center/sanctions/Programs/Documents/cyber2_eo.pdf, at 1(e).

243. U.S. Dep't of the Treasury, Issuance of Amended Executive Order 13694; Cyber-Related Sanctions Designation (Dec. 16, 2016), https://www.treasury.gov/resource-center/sanctions/OFAC-Enforcement/Pages/20161229.aspx

244. 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." (italics added). U.S. Dep't of the Treasury, OFAC FAQs: Other Sanctions Programs, https://www.treasury.gov/resource-center/faqs/Sanctions/Pages/faq_other.aspx.

245. Press Release, U.S. Dep't of the 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.

Electronic copy available at: https://ssrn.com/abstract=3213954

information" and illegal use of stolen personally identifiable information so as to muddy the precedent.[246]

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

Up to this point in our survey of responses to the growing threat posed by proliferating and improving deep-fake technology, we have focused on the incentives that government can create to drive behavior. What about the marketplace?

We anticipate two types of market-based reactions to the deep-threat challenge. 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. The emergence in recent years of an array of services responding to customer anxieties about identity theft and the like provides an example of this. Second, we expect at least some social media companies on their own initiative to take steps 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 a contrary course might produce down the road in terms of legislative interventions. Both prospects offer benefits, but there are both limits and risks as well.

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

246. *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.")

Electronic copy available at: https://ssrn.com/abstract=3213954

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.

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; the service otherwise 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 states criminalize the interception of electronic communications unless all parties to a communication consent to the interception.[247]

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 sizeable advantage.

247. *See* Danielle Keats Citron, *Spying, Inc.*, 72 WASH. & LEE L. REV. 1243, 1262 (2014) (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, 1485. So long as one party to communications consent to interception, the remaining state laws—38—and federal law permit the practice.

Electronic copy available at: https://ssrn.com/abstract=3213954

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.

Should we encourage the emergence of such services, then? We urge caution. 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"[248]—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.[249] 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 supplier of these services. The scale and nature of the data they would host would be 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."[250] The Cambridge Analytica/Facebook scandal might pale in comparison to the possibilities unleashed by such a database.

What's more, the third-party doctrine (at least as things currently stand) would ensure relatively easy government access to that database for investigative purposes. And certainly government investigators would begin to look, first and foremost, towards this rich trove of information in most if not all cases.[251] While it is possible that doctrinal developments at the Supreme Court (in *Carpenter*[252] and its eventual progeny) may yet require the government to seek warrants to obtain such comprehensive data-collections

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

249. *Id.* at 1181.

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

251. Neil M. Richards, *The Third Party Doctrine and the Future of the Cloud,* 94 WASH. U. L. REV. 1441, 1442 (2017).

252. United States v. Carpenter, 819 F.3d 880 (6th Cir. 2016), *cert. granted*, 137 S. Ct. 2211 (2017).

Electronic copy available at: https://ssrn.com/abstract=3213954

held by third parties (or that Congress might do the same via statute as applied to such a new and impactful service industry), the fact remains 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).

This would have its upsides, certainly, in terms of identifying those guilty of crime and avoiding mistaken prosecution of the innocent. But, critically, it also 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 or 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.[253] Again, the resulting enhancement of prosecutorial capacity will be welcome in some quarters, but may cause considerable heartburn in others. At the very least, this warrants careful consideration by policymakers and lawmakers.

Ultimately, a world with widespread lifelogging of this kind might yield more good than harm, particularly if there is legislation well-tailored to regulate access to such a new state of affairs. 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 ages past with the legal architecture for control of traditional public fora tended to loom largest.[254] Today's most important speech fora, for better or worse, are the platforms. And TOS agreements determine if speech on the platforms is visible, prominent, or viewed, or if instead it is hidden, muted, or never available at all.[255] TOS agreements thus will be primary battlegrounds in the fight to minimize the harms that deep fakes may cause.

---

253. Though the third-party doctrine was not actually modified in *United States v. Jones*, 565 U.S. 400 (2012), a majority of the 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. David Gray & Danielle Keats Citron, *The Right to Quantitative Privacy*, 98 MINN. L. REV. 62, 64-65 (2013).

254. Citron & Richards, *supra* note.

255. Klonick, *supra* note, at 1630–38.

Electronic copy available at: https://ssrn.com/abstract=3213954

Some TOS agreements would already ban certain categories of s. For instance, Twitter has long banned impersonation, without regard to the technology involved in making the impersonation persuasive.[256] 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."[257] Technological due process in the first instance 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. Speech policies also should recognize that some deep fakes are not on balance problematic and should remain online; those that constitute satire, parody, art, or education, as explored above, should not normally be suppressed.

Platforms should provide accountability for their speech-suppression decisions, moreover. Users should be notified that their (alleged) deep-fake posts have been removed (or muted) and given a meaningful chance to challenge the decision. 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.[258] 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.[259] It may have a useful 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

256. CITRON, HATE CRIMES IN CYBERSPACE, *supra* note.

257. *Id.* 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, at 1668-69.

258. Citron & Jurecic, *supra* note.

259. *Cf.* Georgia Wells et al., *Russia Bypassed Facebook Filters*, WALL ST. J., Feb. 23, 2018 (reporting that YouTube mistakenly promoted a conspiratorial video falsely accusing a teenage witness to the Parkland school shooting of being an actor).

Electronic copy available at: https://ssrn.com/abstract=3213954

and has stated that it will begin tracking fake videos.[260] Also underway are efforts to emphasize videos from verified sources while also affirmatively deemphasizing ones that do 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.[261] Much more will be needed, but the start is welcome.

IV. 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.

260. Wells, *supra* note, at 18.

261. Wells, *supra* note, at.

Electronic copy available at: https://ssrn.com/abstract=3213954