DISRUPTING DISINFORMATION: DEEPFAKES AND THE LAW

*Anna Yamaoka-Enkerlin**

Current global crises—from health pandemics to election instability—underscore the impact of misinformation, disinformation, and malinformation on society. On March 7, 2020, the N.Y.U. Journal of Legislation & Public Policy *and the NYU Center for Cyber Security hosted a symposium, "When Seeing Isn't Believing: Deepfakes and the Law." This Comment reviews some of the symposium speakers' discussions of the legal issues that the emergence of deepfake technology presents. It also adopts the "Disinformation Disruption Framework" developed by the DeepTrust Alliance to analyze proposed solutions to the local, organizational, and global threats that deepfake technology poses.*

INTRODUCTION

On February 2nd, 2020, the World Health Organization announced that, along with the COVID-19 pandemic, we are also facing an infodemic: "an over-abundance of information—some accurate and some not—that makes it hard for people to find trustworthy sources and reliable guidance when they need it."[1] Exactly one month later, the NYU Center for Cyber Security and the *N.Y.U. Journal of Legisla-*

* Anna Yamaoka-Enkerlin, B.A., University of Oxford; LLM., New York University. Many thanks to Cam Brewer for his feedback during the drafting process, as well as to the editorial team at the *N.Y.U. Journal of Legislation & Public Policy.*

1. World Health Org. [WHO], *Novel Coronavirus (2019-nCov) Situation Report - 13* (Feb. 2, 2020), https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200202-sitrep-13-ncov-v3.pdf [https://perma.cc/A9HB-2BXJ].

tion & Public Policy hosted a symposium, "When Seeing Isn't Believing: Deepfakes and the Law," focusing on the legal and regulatory response to the emergence of one particular source of inaccurate information: deepfakes.

What follows is an attempt to document and process some of the many insights that these panelists—who ranged from tech representatives to policy experts to reporters—shared. First, I will introduce the terms of engagement as they were explained by the event's keynote speaker, Kathryn Harrison, the CEO of DeepTrust Alliance. Next, I will canvas the present threat of deepfakes to the information environment. Finally, I will analyze the range of possible technical, regulatory, and educational solutions raised using a "disinformation misinformation framework."

I.
Terms of Engagement

Deepfake, a combination of 'deep learning' and 'fake,' refers to images, videos, audio or text that is created using AI techniques such as General Adversarial Networks (GANs).[2] GANs are often explained as analogous to a counterfeiter who is learning to make counterfeit money.[3] The counterfeiter's adversary is the treasury, which evaluates notes for authenticity. They are engaged in an arms race—both are constantly learning and improving their respective faking and detection capabilities. In the context of GANs, the counterfeiter is like the 'generator' neural network. The treasury is like another neural network called the 'discriminator.' Like a counterfeiter, the generator's goal is to minimize the probability that the discriminator correctly assigns the label of real or fake to its output. Conversely, the discriminator's objective is to maximize the probability of correctly labeling data as either generated or belonging to a dataset of "real" content. This creates a feedback loop which powers iterative learning. The goal is that with enough training, the generator will improve to the point that

2. *See generally* Ian J. Goodfellow et al., *Generative Adversarial Nets*, 27 Advances Neural Info. Processing Sys. 2672 (2014) [https://perma.cc/Z3TZ-MD74].

3. *See* N.Y.U. School of Law, *Keynote: Kathryn Harrison, Founder & CEO, DeepTrust Alliance*, YouTube (June 30, 2020), https://www.youtube.com/watch?v=sf0Mm4t9kMQ&list=plJkLD_s9pYaZU_FpkX_kmH1jh0wjxgM08 [hereinafter *Keynote*]; Ian J. Goodfellow et al., *Generative Adversarial Nets*, arXiv.org (June 10, 2014), https://arxiv.org/pdf/1406.2661.pdf (introducing deepfake technology and explaining that the technology is like "a team of counterfeiters" that creates *fake* currency that is indistinguishable from *real* currency.).

the probability of the discriminator accurately distinguishing the real from the generated approaches zero.[4]

The result is increasingly realistic voice, text, and audio content. Some deepfakes, like Elon Musk making an appearance in a StarTrek episode, can be purely entertaining.[5] Another example, more alarming in its implications, is a deepfake of Barack Obama appearing to insult Donald Trump.[6] Audio deepfakes, such as clips of Joe Rogan's voice,[7] add additional layers of complexity to increasingly realistic content.[8]

Deepfakes are distinct from "cheapfakes."[9] A cheapfake is a piece of content created using cheap, accessible software (or no software at all) to speed up, slow down, crop, re-contextualize, or otherwise manipulate meaning. For example, researchers have exposed how Alexander Gardner, the famous Civil War photographer, in fact staged many of his most striking images, which he had represented as being candid.[10] A recent widely reported example was a video of Democratic Leader Nancy Pelosi, which was subtly slowed down to make it look like she was slurring her speech.[11]

---

4. *A Beginner's Guide to Generative Adversarial Networks (GANs)*, PATHMIND, http://pathmind.com/wiki/generative-adversarial-network-gan (last visited Mar. 25, 2020) [https://perma.cc/LGB6-3HK3].

5. *See* Think Sink, *Elon Musk as Barclay from Star Trek [Deepfake]*, YOUTUBE (Aug. 11, 2019), https://www.youtube.com/watch?v=kvkIvDR9aGY [https://perma.cc/G39V-GX79].

6. *See* BuzzFeedVideo, *You Won't Believe What Obama Says in This Video!*, YOUTUBE (Apr. 17, 2018), https://www.youtube.com/watch?v=CQ54GDm1eL0 [https://perma.cc/3NR2-BSA3].

7. *See* Dessa, *RealTalk: We Recreated Joe Rogan's Voice Using Artificial Intelligence*, YOUTUBE (May 10, 2019), https://www.youtube.com/watch?v=DWK_iYBl8cA [https://perma.cc/6H3X-ZA7S].

8. During the conference, for example, Harrison played the deepfakes of Joe Rogan's voice and had the audience raise their hands and vote on whether the clip was a deepfake or something Rogan actually said. There was no consensus. *See Keynote*, *supra* note 3.

9. Lawmaker's bills have often conflated the two, resulting in confusion over the scope of proposed legislation. *See* Hayley Tsukayama, India McKinney, & Jamie Williams, *Congress Should Not Rush to Regulate Deepfakes*, ELECTRONIC FRONTIER FOUND. (June 24, 2019), https://www.eff.org/deeplinks/2019/06/congress-should-not-rush-regulate-deepfakes [https://perma.cc/J3PY-3RHF].

10. *The Case of the Moved Body*, LIBR. CONGRESS, https://www.loc.gov/collections/civil-war-glass-negatives/articles-and-essays/does-the-camera-ever-lie/the-case-of-the-moved-body/ (last visited Apr. 20, 2020). Although not usually couched in the language of "cheapfakes," ethical issues around misquoting and staged photographs are a persistent feature of photojournalism. *Staging, Manipulation, and Truth in Photography*, N.Y. TIMES (Oct. 16, 2015), https://lens.blogs.nytimes.com/2015/10/16/staging-manipulation-ethics-photos/.

11. Sarah Mervosh, *Distorted Videos of Nancy Pelosi Spread on Facebook and Twitter, Helped by Trump,* N.Y. TIMES (May 24, 2019), https://www.nytimes.com/2019/05/24/us/politics/pelosi-doctored-video.html [https://perma.cc/J92X-HVD2].

More obviously than deepfakes, cheapfakes defy simple categorizations into 'fake' and 'real.' This may explain why Facebook, long resistant to becoming an "arbiter[ ] of the truth,"[12] introduced a ban against deepfakes on its platform in January 2020.[13] From an effects-based point of view, this distinction seems arbitrary.[14] But having a technological hook on which to hang regulation avoids otherwise difficult decisions that many intermediaries would rather avoid altogether.

Understanding the categories of information is critical to understanding and communicating about the landscape at issue.[15] This can, at minimum, lead to a much more nuanced and productive discussion than one framed in terms of the politicized blanket term "fake news." Though scholars have explicated many different classifications,[16] there are three critical categories to discuss[17]:

- Misinformation: False information that is not created or distributed with the intention to cause harm—e.g., the well-meaning sharing of hoax COVID-19 remedies.
- Malinformation: True information which is used with an intention to harm—e.g., the release of embarrassing intimate photos.
- Disinformation: False information created with an intent to harm or change perceptions of reality and truth—e.g., the "Pizzagate" conspiracy that went viral in advance of the 2016 United States presidential election.[18]

---

12. Adam Mosseri, *Working to Stop Misinformation and False News*, FACEBOOK MEDIA (Apr. 7, 2017), https://www.facebook.com/facebookmedia/blog/working-to-stop-misinformation-and-false-news [https://perma.cc/4YV2-CW5L] ("We cannot become arbiters of truth ourselves—it's not feasible given our scale, and it's not our role.").

13. *See* Monika Bickert, *Enforcing Against Manipulated Media*, FACEBOOK (Jan. 6, 2020), https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/ [https://perma.cc/H6C2-QDWQ].

14. *See* Tony Romm et al., *Facebook Bans Deepfakes, But New Policy May Not Cover Controversial Pelosi Video*, WASH. POST (Jan. 7, 2020, 3:56 PM), https://www.washingtonpost.com/technology/2020/01/06/facebook-ban-deepfakes-sources-say-new-policy-may-not-cover-controversial-pelosi-video/ [https://perma.cc/3CX3-HPYJ].

15. *See Keynote*, *supra* note 3.

16. *See generally* Maria D. Molina et al., *"Fake News" Is Not Simply False Information: A Concept Explication and Taxonomy of Online Content*, AM. BEHAV. SCIENTIST (2019), https://doi.org/10.1177/0002764219878224.

17. *See Keynote*, *supra* note 3.

18. Claire Wardle & Hossein Derakhshan, *Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking* 20, Council of Europe DGI(2017)09 (Sept. 27, 2017), https://rm.coe.int/information-disorder-report-november-2017/1680764666.

II.
THREATS TO THE INFORMATION ENVIRONMENT: "IS THIS A NEW FLAVOR OF A FOREVER PROBLEM?"

As Sun Tzu, the Sixth Century philosopher and military strategist, wrote in *The Art of War*: "All warfare is based on deception."[19] From the Roman empire to present day, there is a long history of actors using disinformation to manipulate public opinion.[20] This begs the question, "Is this a new flavor of a forever problem?"[21]

It is important to remember that this is not the first time that technology has created anxiety around the truth.[22] Although deepfakes are a growing phenomenon (an October 2019 study by DeepTrace labs found that the number of deepfake videos had nearly doubled in the preceding 7 months[23]), cheapfakes and garden variety textual misinformation are by far the greater challenge we presently face. As Ben Wizner, Director of the Speech, Privacy and Technology Project at the ACLU puts it, "focusing on deepfakes is like looking through a straw."[24]

As for claims that deepfakes are "something new" because of their apparently heightened believability, some have expressed skepticism about whether visual deepfakes are really inherently more persuasive than, say, forged documents.[25] Many malicious actors can accomplish their goals without resorting to deepfake technology. This was highlighted in June 2019, when it was revealed that as part of its research into influence campaigns, Jigsaw, an independent Google

19. *See* N.Y.U. School of Law, *The Front Line: Big Tech, Fake News, and Private Industry's Deepfake Detection Problem*, YOUTUBE [hereinafter *The Front Line*] (June 30, 2020), https://www.youtube.com/watch?v=IL-QmxMKcCo&list=PLJkLD_s9pYaZU_FpkX_kmH1jh0wjxgM08&index=2 (Judi Germano, a distinguished fellow at the N.Y.U. Center for Cybersecurity, kicked off the first panel by quoting Sun Tzu).

20. *Id.*; *see* Izabella Kaminska, *A Lesson in Fake News from the Info-Wars of Ancient Rome*, FIN. TIMES (Jan. 17, 2017), https://www.ft.com/content/aaf2bb08-dca2-11e6-86ac-f253db7791c6.

21. *See The Front Line*, *supra* note 19 (as phrased by Germano in her opening question to the panelists).

22. N.Y.U. School of Law, *Legislative Solutions, Individual Rights, and the Question of Government Intervention*, YOUTUBE (June 30, 2020), https://www.youtube.com/watch?v=81Ppe7Vmo8o&list=PLJkLD_s9pYaZU_FpkX_kmH1jh0wjxgM08&index=4 (As noted by Ben Wizner, Director of the Speech, Privacy and Technology Project at the ACLU).

23. Giorgio Patrini, *Mapping the Deepfake Landscape*, DEEPTRACE (July 10, 2019), https://deeptracelabs.com/mapping-the-deepfake-landscape/ [https://perma.cc/8QSC-WCH8].

24. N.Y.U. School of Law, *supra* note 22.

25. *See The Front Line*, *supra* note 19 (as noted by Andrew Gully during the conference).

unit dedicated to developing technological insights on emerging issues, had tested out disinformation-for-hire services from a Russian Troll company—a move that sparked significant controversy.[26] But, as a former prosecutor has replied, based on their experience handling evidence in court[27] deepfakes are a particularly powerful new weapon in the arsenal of malicious actors.

The idea that deepfakes may inflict particular emotional, reputational, and dignitary harms that result from the non-consensual digitization of our bodies and voices is another reason to think that deepfakes really are "something new." Danielle Citron Keats describes how cyberharassment and cyberstalking can be described as Hate 3.0, "because they amount to personalized hate" that manifests in our tailored online experiences.[28] Deepfakes might be the byword of Hate 3.0. And because technologies such as GANs are improving to require less and less training data, this can occur at an unprecedented scale. Few are safe: "With thousands of images of many of us online, in the cloud, and on our devices, anyone with a social media profile is fair game to be faked."[29]

This is especially alarming given the way that women, queer people, and other minorities already disproportionately experience cyber harassment.[30]

III.
The Risks: From Local to Global

There are two main takeaways from the discussion of how deepfakes and other kinds of manipulated media exacerbate the threats posed by disinformation.

First, deepfakes pose a dual threat. The most obvious variety of threat stems from minds being won over by the disinformation itself.

---

26. Andy Greenberg, *Alphabet-Owned Jigsaw Bought a Russian Troll Campaign as an Experiment*, Wired (June 12, 2019, 10:12 AM), https://www.wired.com/story/jigsaw-russia-disinformation-social-media-stalin-alphabet/ [https://perma.cc/2YQN-GLKR].

27. *See The Front Line*, *supra* note 19 (Andrew Gully, the Technical Research Manager at Jigsaw, expressed this skepticism during the conference over whether deepfakes will really change the landscape for malicious actors. In reply to Gully's skepticism, Germano described her experience as a former prosecutor.).

28. Danielle Keats Citron, Hate Crimes in Cyberspace (Harvard Univ. Press 2014).

29. Britt Paris & Joan Donovan, Data & Soc'y, Deepfakes and Cheapfakes: The Manipulation of Audio and Visual Evidence 7 (2019), https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1.pdf. [https://perma.cc/9H3H-8APZ].

30. Citron, *supra* note 28, at 14–16.

The other is the "Liars Dividend"—the phenomenon that arises when the ability to create convincing fakes allows actors to undermine the veracity of real, accurate information by claiming that it, too, is fake.[31] As an example of the liar's dividend at work,[32] in 2018, President Ali Bongo Ondimba of Gabon had spent several months out of the country, apparently seeking medical treatment.[33] In response to rumors that he was dead or incapacitated, on New Year's Day the president released a 'proof of life' video to show that he was recovering from a stroke.[34] Days later, the president's opponents claimed that this video was a deepfake and attempted a coup.[35]

The second take-away is that every level of society, from the individual to the global community, is at an increased risk of harm.

At the individual level, the most novel threat posed by deepfakes is the ability to create realistic, non-consensual pornography and other intimate images that can be used to embarrass, harass, and extort. Pornography, mostly featuring the faces of female celebrities superimposed onto other bodies, accounts for 96% of online deepfake content available today.[36] Corin Faife, a journalist and researcher working on the Emerging Threats and Opportunities program at WITNESS, is particularly concerned about a tendency to over-focus on the theoretical possibility of deepfake-induced geopolitical instability at the expense of tackling the present threat posed by the weaponization of deepfake pornography.[37]

The experience of Rana Ayyub illustrates this threat. Ayyub, an Indian journalist and outspoken critic of the Hindu Nationalist movement, was the target of a deepfake porn plot. She describes what happened when someone forwarded her the video over WhatsApp. "I started throwing up . . . I just started crying. It was devastating. I just couldn't show my face."[38] As the video spread across India through

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

32. *See Keynote*, *supra* note 3 (using the recent events in Gabon as an example of the liar's dividend).

33. *See* Sarah Cahlan, *How Misinformation Helped Spark an Attempted Coup in Gabon*, WASH. POST (Feb. 13, 2020, 3:00 AM), https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/.

34. *See id.*

35. *See id.*

36. Patrini, *supra* note 23.

37. *The Front Line*, *supra* note 19.

38. Rana Ayyub, *I Was the Victim of a Deepfake Porn Plot Intended To Silence Me*, HUFFPOST UK (Nov. 21, 2018, 8:11 AM), https://www.huffingtonpost.co.uk/entry/deepfakeporn_uk_5bf2c126e4b0f32bd58ba316.

Facebook and WhatsApp, Ayyub was the subject of online abuse and death threats. Eventually, the United Nations had to intervene. Though this is a relatively extreme example, there need not be widespread distribution for there to be an impact on one's reputation. Indeed, there need not be any distribution at all—just the threat of a video's release may be enough for a blackmailer to achieve their ends.

At the institutional level, deepfakes may pose new challenges in the courtroom. Given that one of the core functions of our judicial system is to adjudicate on facts and discern the truth, deepfakes raise obvious evidentiary challenges.[39] One can easily imagine an altered surveillance video or voice recording that is realistic enough to convince a jury to wrongfully convict.

Deepfakes also pose threats to businesses, as well as entire economies. Scammers have been training programs to carry out sophisticated fraud, including by impersonating CEOs using algorithms trained on snippets of speech from earnings calls, YouTube videos, and TED Talks.[40] This tactic has already been used to steal from individual corporations, but given that statements by influential CEOs or other shocking news stories often have immediate impacts on stock price, we might see deepfakes being used to manipulate the stock market.[41]

Emerita Torres, Director of Policy Research and Programs at the Soufan Centre, has discussed how misinformation and disinformation makes it more challenging for government policy to be carried out.[42] This has been experienced acutely in the wake of COVID-19—the

39. *See also* Kathryn Lehman et al., *5 Ways to Confront Potential Deepfake Evidence in Court*, LAW360 (July 26, 2019, 4:59 PM), https://www.law360.com/articles/1181306/5-ways-to-confront-potential-deepfake-evidence-in-court.

40. Kaveh Waddell & Jennifer A. Kingson, *The Coming Deepfakes Threat to Business*, AXIOS (July 19, 2019), https://www.axios.com/the-coming-deepfakes-threat-to-businesses-308432e8-f1d8-465e-b628-07498a7c1e2a.html [https://perma.cc/7JRK-TBSY]; Catherine Stupp, *Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case*, WALL STREET J. (Aug. 30, 2019, 12:52 PM), https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402 [https://perma.cc/MW9H-A474].

41. In 2013, Syrian hackers claimed credit for hacking AP News' Twitter account and tweeting that there had been an explosion in the White House and that President Obama was injured. $136 billion was momentarily wiped from the stock market. Max Fisher, *Syrian Hackers Claim AP Hack That Tipped Stock Market by $136 Billion. Is it Terrorism?*, WASH. POST (Apr. 23, 2013, 4:31 PM), https://www.washingtonpost.com/news/worldviews/wp/2013/04/23/syrian-hackers-claim-aphack-that-tipped-stock-market-by-136-billion-is-it-terrorism [https://perma.cc/KBC2-2XHH].

42. *See* N.Y.U. School of Law, *Global Implications of False Information for National and International Security and Human Rights*, YOUTUBE [hereinafter *Global Implications*] (June 30, 2020), https://www.youtube.com/watch?v=GY6AXMyb-yY&list=PLJkLD_s9pYaZU_FpkX_kmH1jh0wjxgM08&index=3.

"perfect storm for conspiracy theorists"—which has seen protestors congregate in defiance of stay-at-home orders.[43] "If the one in three Americans who believes that the effects of COVID-19 have been exaggerated choose to forgo crucial health practices, such as social distancing . . . the disease could spread faster and farther than otherwise, and could cost many thousands of lives."[44]

With the 2020 election approaching, the destabilizing effect of misinformation on political discourse and the specter of election interference also loomed large in the panelists' conversation. The biggest threat to the 2020 election, Gorman said, would be a perfectly timed, politically compromising video released just before election day, when there is not enough time to verify the video or undo the damage.

Finally, the conference featured a panel focused on the impact of deepfakes at the global level, with a focus on terrorism and other geopolitical implications. Imagine a video showing the Israeli prime minister apparently involved in planning an assassination in Iran, or depicting an American general burning a copy of the Koran.[45] As Torres underscored, in the context of societies already riven by cultural and political fissures, the potential for these recordings to incite violence is especially great.

At the same time, Mounir Ibrahim, Vice President of Strategic Initiatives at Truepic, has discussed how public-private partnerships could incorporate deepfakes into "industrial level misinformation campaigns," and how the destruction of the information environment within democracies plays into authoritarian hands.[46] As a former U.S. foreign service officer and key advisor on Syria, he has addressed the challenges that arise when multilateral forums, such as the UN Security Council, are expected to make decisions based on user-generated content from exclusion zones. "I saw the liars dividend play out . . . in political decision making. If you question the veracity of images, you could undermine the situation on the ground and stall international action."[47]

---

43. Joseph E. Uscinski & Adam M. Enders, *The Coronavirus Conspiracy Boom*, ATLANTIC (Apr. 30, 2020), https://www.theatlantic.com/health/archive/2020/04/what-can-coronavirus-tell-us-about-conspiracy-theories/610894/.

44. *Id.*

45. Robert Chesney & Danielle Citron, *Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics*, FOREIGN AFFAIRS (Dec. 11, 2018, 12:00 AM), https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war [https://perma.cc/5GGN-89DB].

46. *See Global Implications*, *supra* note 42.

47. *Id.*

IV.
Framing the Solutions

The "Disinformation Disruption Framework" developed by the DeepTrust Alliance breaks down the disinformation landscape into five phases: Makers, Creation, Distribution, Believability, and Impact. Disaggregating disinformation into these phases highlights the unique challenges and opportunities that each phase presents, which helps to identify key stakeholders and evaluate interventions.[48]

There are different avenues to resolve unique challenges of disrupting disinformation, such as technological development (e.g. detection technology), public regulation (e.g. legislation), private regulation (e.g. terms of service), and public education. Specific proposals from each of these groups can be plotted on every level of the disinformation disruption framework.

A. Makers

The Makers phase refers to the individuals and organizations that decide to create content. This phase reminds us that it's not just nefarious actors that see potential in deepfakes.

For example, artists and educators are paying attention to "Dali Lives," an exhibition at the Dali Museum in St. Petersburg, featuring a lifelike interactive Salvador Dali that was created by training an algorithm using archival footage.[49] Filmmakers see deepfakes as an improvement over current CGI techniques,[50] though the ethics of synthetic resurrection remain murky and raise questions about consent and the commercial exploitation of our digital afterlives.[51]

GANs are also being used to compensate for a shortage of quality data sets for training algorithms. Despite the allure of "big data," for

48. Kathryn Harrison & Sara Aros, DeepTrust All., Deepfake, Cheapfake: The Internet's Next Earthquake? (2020), https://static1.squarespace.com/static/5d894b6dcd6a2255c38759fe/t/5e44d9257a6edf3b61208568/1581570371567/DeepTrust+Report+1.

49. Dami Lee, *Deepfake Salvador Dalí Takes Selfies with Museum Visitors*, Verge (May 10, 2019, 8:50 AM), https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum [https://perma.cc/J7S7-VWNG].

50. Erin Winick, *How Acting as Carrie Fisher's Puppet Made a Career for Rogue One's Princess Leia*, MIT Tech. Rev. (Oct. 16, 2018), https://www.technologyreview.com/s/612241/how-acting-as-carrie-fishers-puppet-made-a-career-for-rogue-ones-princess-leia [https://perma.cc/E4Q5-MM4W].

51. Carl Öhman & Luciano Floridi, *An Ethical Framework for the Digital Afterlife Industry*, 2 Nature Hum. Behav., 318, 318–20 (2018). In New York, a now-expired bill proposed extending protection of one's portrait—including one's digitally manipulated likeness—for 40 years after death. Assemb. B. 8155, 2017-2018 Leg. Sess. (N.Y. 2017).

data to be useful it's not enough to simply have a lot of it. Though there is no universal definition of quality data, common indicators include accessibility, processability and cleanliness.[52] Accessibility of health data can be especially challenging because of mounting privacy concerns. To overcome this issue, researchers are using GANs to create datasets of 'fake' brain scans. Algorithms aimed at spotting tumors became just as good as algorithms trained only on real images when trained on a data set made up of only 10% real scans.[53]

Turning to the malicious makers of deepfakes, threat-attribution work can identify and disable vectors of deception, from lone wolves to nation-state threats.[54]

Resolving the threat-attribution problem is the first barrier to minimizing the reach of harmful makers.[55] In August 2019, Twitter announced that media "financially or editorially controlled by the state" would be prevented from using its advertising services.[56] The announcement was made after investigations revealed that Chinese-controlled media companies were using Twitter to push disinformation about protests in Hong Kong. But this move likely only escalates the cat-and-mouse attribution game as states and other banned actors develop increasingly sophisticated techniques to conceal their influence over apparently independent proxies.[57]

The problem is exacerbated by the fact that in the private sector, most terms of service and community guidelines do not support detec-

52. *Open Data Quality–the Next Shift in Open Data?*, OPEN KNOWLEDGE FOUND. (May 31, 2017), https://blog.okfn.org/2017/05/31/open-data-quality-the-next-shift-in-open-data/ [https://perma.cc/6K56-D9S8].

53. Hoo-Chang Shin et al., *Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks*, *in* THIRD INTERNATIONAL WORKSHOP ON SIMULATION AND SYNTHESIS IN MEDICAL IMAGING, 11037 LECTURE NOTES IN COMPUTER SCIENCE 1, 8 (Ali Gooya et al. eds. 2018), http://arxiv.org/abs/1807.10225.

54. *See* N.Y.U. School of Law, *Legislative Solutions, Individual Rights, and the Question of Government Intervention*, YOUTUBE (June 30, 2020), https://www.youtube.com/watch?v=81Ppe7Vmo8o&list=PLJkLD_s9pYaZU_FpkX_kmH1jh0wjxgM08&index=4 (as discussed by Robert Volkert, Vice President of Threat Investigations at the Cybersecurity firm Nisos).

55. W. Earl Boebert, *A Survey of Challenges in Attribution*, *in* PROCEEDINGS OF A WORKSHOP ON DETERRING CYBERATTACKS: INFORMING STRATEGIES AND DEVELOPING OPTIONS FOR U.S. POLICY (2010).

56. *Updating Our Advertising Policies on State Media*, TWITTER: BLOG (Aug. 19, 2019), https://blog.twitter.com/en_us/topics/company/2019/advertising_policies_on_state_media.html.

57. Camille François, Transatlantic Working Grp., Actors, Behaviors, Content: A Disinformation ABC at 3 (Sept. 20, 2019) (unpublished manuscript), https://www.ivir.nl/publicaties/download/ABC_Framework_2019_Sept_2019.pdf [https://perma.cc/KL7J-DHFU].

tion and enforcement against harmful actors. There is also a dearth of international cybersecurity cooperation.[58] Proposals to enhance global cooperation include updating the Budapest Convention of Cybercrime and updating internet protocols to make attribution more effective on the technical level. "This will help verify compliance with principles of international law such as noninterference in the internal affairs of other states—including elections—and hold states more responsible for what happens in their cyber realm."[59]

As the attribution problem worsens, experts have warned that "attribution fixation" may lead to a paralysis of action[60]—suggesting that intervention at a different phase of the disinformation framework may be more effective.

B. Creation

This phase is about the creation of deepfakes through hardware, software, and human input.[61]

How easy is it to create a deepfake? Creating a widely convincing deepfake is currently an arduous feat.[62] But if the goal is to humiliate or sow doubt, then deepfakes that are far less than perfect may do the trick. And the technology to create them is becoming increasingly accessible.

There are three main approaches to creating deepfakes.[63] The first approach is to use open source tools. Under open source software licenses, authors can make software accessible and grant users the right to copy, modify, redistribute, and use the software for any purpose.[64] GitHub, an open source development platform now owned by Microsoft, hosts most deepfake repositories. The most active focuses

58. Elena Chernenko, Oleg Demidov & Fyodor Lukyanov, *Increasing International Cooperation in Cybersecurity and Adapting Cyber Norms*, COUNCIL ON FOREIGN REL. (Feb. 23, 2018), https://www.cfr.org/report/increasing-international-cooperation-cybersecurity-and-adapting-cyber-norms [https://perma.cc/G5QR-5MCY].

59. *Id.*

60. JASON HEALEY, ATL. COUNCIL, BEYOND ATTRIBUTION: SEEKING NATIONAL RESPONSIBILITY FOR CYBER ATTACKS (2012), https://www.atlanticcouncil.org/wp-content/uploads/2012/02/022212_ACUS_NatlResponsibilityCyber.PDF [https://perma.cc/2A3Q-NR8S].

61. *Id.*

62. *See* N.Y.U. School of Law, *The Front Line: Big Tech, Fake News, and Private Industry's Deepfake Detection Problem*, YOUTUBE (June 30, 2020), https://www.youtube.com/watch?v=IL-QmxMKcCo&list=PLJkLD_s9pYaZU_FpkX_kmH1jh0wjxgM08&index=2 (according to Gully).

63. Robert Volkert & Henry Ajder, *Analyzing the Commoditization of Deepfakes*, N.Y.U. J. LEGIS. & PUB. POL'Y QUORUM (2020).

64. *Licenses & Standards*, OPEN SOURCE INITIATIVE, https://opensource.org/licenses (last visited Apr. 21, 2020) [https://perma.cc/B95T-HZ9D].

on faceswapping and promises "deepfake software for all."[65] Contributors constantly update these repositories to incorporate the latest advances. These repositories are also accompanied by detailed tutorials and active discussion groups of researchers, hobbyists and others on platforms like Discord, Reddit, and Voat.[66]

Faife explained that though the average citizen doesn't have the programming background and powerful graphics processor needed to harness these resources, prospective makers could easily commission a creator. 'Deepfakes-as-a-service' is a nascent market, with marketplace sellers mostly advertising in online forums and marketplaces.[67] Though many sellers state that they will not make pornographic or malicious content, others promote their ability to create custom deepfake pornography.[68]

The final approach to creation is through service platforms, which allow users to upload data such as photos and automate the process of creating deepfakes through a user interface. Examples include apps like Faceapp, Zao, and Deepnude. Social media giants Snapchat and TikTok may soon join them.[69]

What sorts of interventions might be considered at the creation level?

Until now, self-regulation has ruled. Reddit voluntarily shut down the original deepfakes SubReddit in February 2018. In July 2019 GitHub cited its terms of service which bar sexual obscenity and took down repositories related to the app 'Deepnude' after the app's developers uploaded its source code onto the platform.[70] Concerned about unethical uses of their code, some individual developers are promoting a shift away from the established open source norms which guarantee freedom-from-use restrictions.[71] For example, to protest the

65. *Deepfakes/Faceswap*, GITHUB, https://github.com/deepfakes/faceswap (last visited Apr. 21, 2020) [https://perma.cc/ERT5-J89N].

66. HEALEY, *supra* note 60.

67. *Id.*

68. *Id.*

69. Michael Nuñez, *Snapchat and TikTok Embrace 'Deepfake' Video Technology Even as Facebook Shuns It*, FORBES (Jan. 8, 2020, 6:30 AM), https://www.forbes.com/sites/mnunez/2020/01/08/snapchat-and-tiktok-embrace-deepfake-video-technology-even-as-facebook-shuns-it/.

70. Joseph Cox, *GitHub Removed Open Source Versions of DeepNude*, VICE (July 9, 2019, 11:26 AM), https://www.vice.com/en_us/article/8xzjpk/github-removed-open-source-versions-of-deepnude-app-deepfakes.

71. Under the Open Source Definition (OSD) promulgated by the Open Source Initiative (OSI), one of the criteria that the distribution terms of open-source software must comply with is "No discrimination against fields of endeavor." Under this term, "The license must not restrict anyone from making use of the program in a specific field of endeavor. For example, it may not restrict the program from being used in a

separation of families at the US-Mexico border, Seth Vargo deleted his code off of GitHub when he found out that a company was relying on his code to fulfill a contract with ICE.[72] Meanwhile, Carolina Ada Ehmke has created a 'Hippocratic License' that can be used in place of the established open source licenses that reject any use restrictions on code.[73]

Another type of self-regulation comes in the form of developing community standards. The deepfakes repository on GitHub itself includes a 'Manifesto' which states: "Are there some out there doing horrible things with similar software? Yes. And because of this, the developers have been following strict ethical standards. Many of us don't even use it to create videos, we just tinker with the code to see what it does. . . . Like any technology, it can be used for good or it can be abused."[74]

The discussion of technology as neutral is "deeply embedded in the originating philosophy of open source."[75] But some tools have more potential for abuse than others. There is no doubt that the code GitHub hosts enables people to create devastatingly harmful content. Manifesto aside, we know that 96% of deepfake content, much of which was enabled by this repository, consists of non-consensual pornography.[76]

Since buying GitHub in 2018, Microsoft has failed to openly confront questions about accountability, ethics, and code moderation on GitHub. It is ironic then that Microsoft, in partnership with other com-

---

business, or from being used for genetic research." *The Open Source Definition*, OPEN SOURCE INITIATIVE, https://opensource.org/docs/osd (last updated Mar. 22, 2007) [https://perma.cc/AM9G-DBE3].

72. Zoe Schiffer, *To Fight 'Evil' ICE, an Engineer Pulled His Code off GitHub*, VERGE (Sept. 20, 2019, 7:54 PM), https://www.theverge.com/2019/9/20/20876495/github-seth-vargo-pulled-code-chef-ice-deportations-trump-administration.

73. Klint Finley, *An Open Source License that Requires Users to Do No Harm*, WIRED (Oct. 14, 2019, 8:00 AM), https://www.wired.com/story/open-source-license-requires-users-do-no-harm/.

74. *Deepfakes/Faceswap*, *supra* note 65.

75. Rachel Winter & Anastasia Salter, *DeepFakes: Uncovering Hardcore Open Source on GitHub*, PORN STUD. (Oct. 17, 2019), https://doi.org/10.1080/23268743.2019.1642794 at 7.

76. Davey Winder, *Forget Fake News, Deepfake Videos Are Really All About Non-Consensual Porn*, FORBES (Oct. 8, 2019, 8:35 AM), https://www.forbes.com/sites/daveywinder/2019/10/08/forget-2020-election-fake-news-deepfake-videos-are-all-about-the-porn/#4c711e7563f9.

panies and universities, is spearheading the Deepfakes Detection Challenge, which aims to stop deepfakes' spread.[77]

Despite the shortcomings of self-regulation, several panelists urged against forcing GitHub/Microsoft to take down repositories or otherwise attempt to control open source code. Taking down a piece of offensive content is "not the same as taking down the code that could maybe generate a million pieces of content . . . the implications are different."[78] Rather than moderating code, state attorneys general and other regulators could implement controls on emerging end products,[79] as the New York Attorney General has done in the context of stalkerware.[80]

But some argue that new legislation is unnecessary because creators of harmful deepfakes could already be liable under a panoply of current laws.[81] Apart from criminal laws against extortion and harassment, available tort actions may include false-light invasion of privacy, defamation, intentional infliction of emotional distress, the right of publicity, as well as copyright claims.[82]

Nevertheless, federal and state lawmakers alike are considering legislation aimed at the creators of deepfakes.[83] When assessing this

77. Mike Schroepfer, *Creating a Data Set and a Challenge for Deepfakes*, FACEBOOK (Sept. 5, 2019) https://ai.facebook.com/blog/deepfake-detection-challenge/ [https://perma.cc/4RKZ-6T4P].

78. Louise Matsakis, *How Will Microsoft Handle GitHub's Controversial Code?*, WIRED (June 5, 2018, 3:42 PM), https://www.wired.com/story/microsoft-github-code-moderation/.

79. *See* N.Y.U. School of Law, *Legislative Solutions, Individual Rights, and the Question of Government Intervention*, YOUTUBE (June 30, 2020), https://www.youtube.com/watch?v=81Ppe7Vmo8o&list=PLJkLD_s9pYaZU_FpkX_kmH1jh0wjxgM08&index=4 (as noted by Noah Stein, an Assistant Attorney General in the Bureau of Internet & Technology at the New York State Attorney General's Office).

80. In October 2019, the FTC took the unprecedented step of barring the developers of three stalkerware apps from selling their apps unless steps were taken to ensure that the apps were used for legitimate purposes and that data collected was deleted. Press Release, Fed. Trade Comm'n, FTC Brings First Case Against Developers of "Stalking" Apps (Oct. 22, 2019), https://www.ftc.gov/news-events/press-releases/2019/10/ftc-brings-first-case-against-developers-stalking-apps.

81. *See* N.Y.U. School of Law, *Legislative Solutions, Individual Rights, and the Question of Government Intervention*, YOUTUBE (June 30, 2020), https://www.youtube.com/watch?v=81Ppe7Vmo8o&list=PLJkLD_s9pYaZU_FpkX_kmH1jh0wjxgM08&index=4 (as Wizner argued during the conference).

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

83. MATTHEW F. FERRARO, WILMERHALE, DEEPFAKE LEGISLATION: A NATIONWIDE SURVEY—STATE AND FEDERAL LAWMAKERS CONSIDER LEGISLATION TO REGULATE MANIPULATED MEDIA (2019), https://www.wilmerhale.com/en/insights/client-alerts/20190925-deepfake-legislation-a-nationwide-survey (follow "Deepfake Legislation: A Nationwide Survey").

phenomenon, it is important to be vigilant, lest the government drift into "ministry of truth" territory.[84] As countries around the world enact laws against disinformation, we are already seeing them being used as a front to fine, censor, arrest, and imprison dissidents.[85]

Legislation in the US has so far been relatively narrow in scope. For example, Texas outlawed the creation of deepfake videos of political candidates intended to injure the candidate or influence an election,[86] while Virginia has banned the use of deepfake technology to produce pornography.[87]

The DEEPFAKES Accountability Act, a bill proposed in Congress, takes a different approach to tackle the issue.[88] Under the H.R. 3230, creators of content falling within the bill's scope would be required to watermark and provide written or audio disclosure that their work contains altered visual or audio elements. Also, any "manufacturer of software" who produces software that the manufacturer reasonably believes will be used to create deepfakes must ensure that the software includes the technical capability to insert watermarks and require users to acknowledge their obligation to include watermarks or other disclosures in their creations. Violating these provisions could result in criminal or civil penalties.

Even if the bill overcomes constitutional objections over compelled speech,[89] the attribution problem may make tracking down creators impossible. Arguably the bill's purpose—"to combat the spread of disinformation through restrictions on deep-fake video alteration technology"[90]—could be better achieved at the distribution phase of the disinformation framework.

---

84. *See* N.Y.U. School of Law, *Legislative Solutions, Individual Rights, and the Question of Government Intervention*, YOUTUBE (June 30, 2020), https://www.youtube.com/watch?v=81Ppe7Vmo8o&list=PLJkLD_s9pYaZU_FpkX_kmH1jh0wjxgM08&index=4 (as Lindsay Gorman, Fellow for Emerging Technologies at the Alliance for Securing Democracy, noted).

85. To download a global database of government actions (updated weekly), see *A Guide to Anti-Misinformation Actions Around the World,* POYNTER, https://www.poynter.org/ifcn/anti-misinformation-actions/ (last visited Apr. 2, 2020).

86. S.B. 751, 86th Leg. (Tex. 2019). California and the United States Congress are considering similar bills. *See* Assemb. B. 1280, 2019-20 Reg. Sess. (Cal. 2019); DEEPFAKES Accountability Act, H.R. 3230, 116th Cong. (2019); *see also* FERRARO, *supra* note 83.

87. VA. CODE ANN. § 18.2-386.2 (2017).

88. Though not discussed in depth at the conference, this proposed piece of federal legislation was the subject of the *N.Y.U. Journal of Legislation and Public Policy* Legislation Competition that happened in advance of the symposium.

89. *See, e.g.*, Eugene Volokh, *The Law of Compelled Speech*, 97 TEX. L. REV. 355, 359–61 (2018).

90. DEEPFAKES Accountability Act, H.R. 3230, 116th Cong. (2019).

C. Distribution

The distribution layer refers to the broad dissemination of content via social media platforms. Though disinformation, misinformation, and malinformation have always been a problem, our cyber connectivity is such that information can travel through vast networks quicker than ever before. In this medium, falsehoods have a clear advantage—studies have shown that misinformation reaches more people and spreads faster than the truth.[91] Untangling the reasons for this—from organic reach, paid content boosting, microtargeting, algorithmic newsfeeds, to bots—is complicated. It does not help that billions of people receive their news through ad-based business models that are optimized for engagement as opposed to other norms.[92]

In the early days of cyberspace, digital platforms were presumed to be mere conduits of content.[93] But "Web 2.0" intermediaries[94] are not mere transmitters of communication, but master architects who design their private digital infrastructure in ways which allow them to control access, shape content, and affect user agency.[95] In these circumstances, there is a growing feeling that the entities who are in the best position to respond should be required to take some responsibility for the negative externalities that they profit from and even encourage.

Facebook and other major social networks are responding with different solutions. According to Saleena Khanum Salahuddin, the Cybersecurity Policy Lead at Facebook, there is no quick fix—"pursuing varied solutions is how we create challenges for adversaries."[96] Measures include partnering with fact-checking organizations, disrupting economic incentives (for example, by cutting off advertising privileges for purveyors of disinformation), making reporting easier, adjusting ranking algorithms, improving content moderation algo-

91. Soroush Vosoughi et al., *The Spread of True and False News Online*, 359 SCIENCE 1146, 1147 (2018).

92. *See* N.Y.U. School of Law, *Legislative Solutions, Individual Rights, and the Question of Government Intervention*, YOUTUBE (June 30, 2020), https://www.youtube.com/watch?v=81Ppe7Vmo8o&list=PLJkLD_s9pYaZU_FpkX_kmH1jh0wjxgM08&index=4 (as Wizner pointed out).

93. Tarleton Gillespie, *The Politics of 'Platforms,'* 12 NEW MEDIA & SOC'Y 347, 348 (2010).

94. Tim O'Reilly, *What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software*, 65 COMM. & STRATEGIES 17, at 36-37 (2007).

95. Jean-Christophe Planti & Aswin Punathambekar, *Digital Media Infrastructures: Pipes, Platforms, and Politics*, 41 MEDIA, CULTURE & SOC'Y 163, 166, 171 (2018).

96. N.Y.U. School of Law, *The Front Line: Big Tech, Fake News, and Private Industry's Deepfake Detection Problem*, YOUTUBE (June 30, 2020), https://www.youtube.com/watch?v=iL-QmxMKcCo.

rithms, improving the capability to detect and remove accounts run by bots, and investing in deepfake detection technology.[97]

Some social media companies have also been changing their terms and conditions in an effort to curb the spread of deepfakes. For example, Twitter recently announced its new policy on "synthetic and manipulated media." Unlike Facebook's deepfakes ban, which has been criticized for being too narrow, Twitter's policy applies to 'cheapfakes' as well. If content is deemed to fall within the scope of the policy, Twitter will also consider whether the content is "shared in a deceptive manner" and whether the content is "likely to impact public safety or cause serious harm."[98] Twitter applied its policy for the first time on a video of presidential candidate Joe Biden posted by a member of President Trump's campaign team. The video was clipped to cut off the end of Biden's sentence, so that it sounded like he was endorsing Trump. Twitter added a "manipulated media" label to it—18 hours after the tweet was posted, at which point 5 million people had seen it.[99] After refusing to flag an earlier video that "misrepresented the order of events" in which Speaker of the House Pelosi ripped up Trump's State of the Union address, Facebook eventually decided to flag the video of Biden as "partly false."[100]

Private messaging services can also be unreliable, and even dangerous, information channels.[101] A key privacy feature of services like WhatsApp and Apple's iMessage is end-to-end encryption. Neither the platform nor anyone who is not added to the group can read a chat's contents. But this also creates impenetrable echo-chambers that are out of reach to fact-checking journalists, manipulated media warnings, or algorithms that track the origins of messages to identify malicious actors. WhatsApp, which has over two billion users, has taken steps to reduce the spread of harmful information, such as tagging messages as 'forwarded,' limiting the number of groups a message can be forwarded to at one time to five, and limiting group membership to

97. Mosseri, *supra* note 12.

98. *Synthetic and Manipulated Media Policy*, TWITTER: HELP CTR., https://help.twitter.com/en/rules-and-policies/manipulated-media (last visited Mar. 20, 2020).

99. *Twitter Labels Edited Biden Video 'Manipulated Tweet*,*'* BBC NEWS (Mar. 10, 2020), https://www.bbc.com/news/world-us-canada-51799366.

100. Taylor Hatmaker, *Facebook Flags Biden Video from Trump's Social Media Director as 'Partly False*,*'* TECHCRUNCH (Mar. 9, 2020, 2:35 PM), https://techcrunch.com/2020/03/09/facebook-biden-video-twitter-trump/.

101. Krishna Pokharel & Rojesh Roy, *India Says Rumors About Child Snatching on WhatsApp Led to Mob Killings*, WALL STREET J. (July 5, 2018, 2:25 AM), https://www.wsj.com/articles/india-admonishes-whatsapp-after-deaths-1530730096 [https://perma.cc/EKJ6-5BBP].

256.[102] WhatsApp has also been setting up business accounts run by fact-checking organizations to which users can forward suspicious messages. But the mass spread of COVID-19 misinformation over WhatsApp shows that these technical restraints have not been enough to overcome the problem.[103]

Apart from the most high-profile social media companies, there are millions of online platforms that continue to host, encourage, and profit from user upload of unlawful content, often leaving victims with little recourse. Part of the reason there is no recourse for victims is because of section 230 of the Communications Decency Act (CDA).

Enacted in 1996, section 230 of the CDA grants immunity to providers of "interactive computer service[s]" from liability for user-generated content that is posted on their sites. In *Zeran v. American Online, Inc.*[104] the Fourth Circuit recognized that Congress granted intermediaries immunity under section 230 for two purposes: as a Good Samaritan provision to "encourage interactive computer services and users of such services to self-police the Internet for obscenity and other offensive material"[105] and to "encourage the unfettered and unregulated development of free speech on the Internet."[106]

Some argue against amending section 230 because if platforms are subject to liability for user-generated content, either 1) they won't exist or 2) they will have to assume the role of speech police, engaging in collateral censorship.[107] Collateral censorship occurs when, in order to avoid liability being imposed on themselves, platforms suppress speech, including beneficial speech.[108] Reliance on algorithmic content moderators—which is necessary given the amount and variety

102. Sinduja Rangarajan, *WhatsApp Is a Petri Dish of Coronavirus Misinformation*, MOTHER JONES (Mar. 20, 2020), https://www.motherjones.com/media/2020/03/whatsapp-coronavirus-misinformation/ [https://perma.cc/E4SQ-NDTE].

103. Nicole Nguyen, *In the Coronavirus 'Infodemic,' Here's How to Avoid Bad Information*, WALL STREET J. (Mar. 22, 2020, 9:00 AM), https://www.wsj.com/articles/in-the-coronavirus-infodemic-you-can-manage-the-deluge-of-news-11584882002.

104. Zeran v. Am. Online, Inc., 129 F.3d 327, 330 (4th Cir. 1997).

105. Batzel v. Smith, 333 F.3d 1018, 1028 (9th Cir. 2003) (first citing 47 U.S.C. § 230(b)(4) (2012); then citing 141 CONG. REC. H8469–70 (daily ed. Aug. 4, 1995) (statements of Reps. Cox, Wyden, and Barton); then citing *Zeran*, 129 F.3d at 331; and then citing Blumenthal v. Drudge, 992 F. Supp. 44, 52 (D.D.C. 1998)).

106. *Batzel*, 333 F.3d at 1027–28 (citing *Zeran*, 129 F.3d at 330).

107. *See* N.Y.U. School of Law, *Legislative Solutions, Individual Rights, and the Question of Government Intervention*, YOUTUBE (June 30, 2020), https://www.youtube.com/watch?v=81Ppe7Vmo8o&list=PLJkLD_s9pYaZU_FpkX_kmH1jh0wjxgM08&index=4 (noted by Wizner).

108. Felix T. Wu, *Collateral Censorship and the Limits of Intermediary Immunity*, 87 NOTRE DAME L. REV. 293, 300 (2013).

of content that some platforms host[109]—makes collateral censorship even more likely.[110] Per Wizner: "The normative answer is that you can't have the kind of internet we have with open speech on free forums, and hold platforms responsible."

But others in the field are proposing alternatives to section 230, and arguably they do not all threaten the vibrancy of internet in the way that Wizner suggests. In short, apart from outright revocation[111], alternatives include a category-based approach[112] and preserving section 230 immunity, but only for smaller platforms (e.g., those making below $100 million in revenue).[113] Another proposal that has garnered traction is making immunity conditional on a duty to "take[ ] 'reasonable steps' to ensure that its platform is not being used for illegal ends."[114]

But an alternative could be lobbying for transparency requirements.[115] At the moment, there is little public knowledge as to how proprietary content-moderation algorithms are making determinations about what content is blocked, pushed, or prioritized on the platforms. The Constitution may constrain the government, but who watches the platforms, our new governors?[116] While companies may understandably resist disclosing their source code to the public and their competitors, perhaps an analogy can be made with auditing, where a group of

109. Stuart Macdonald et al., *Regulating Terrorist Content on Social Media: Automation and the Rule of Law*, 15 INT. J. L. CONTEXT 183 (2019) (". . . every minute, 350,000 tweets are posted, 300 hours of video are uploaded to YouTube and, on Facebook, 510,000 comments are posted, 293,000 statuses are updated and 136,000 photos are uploaded.").

110. Maayan Perel & Niva Elkin-Koren, *Accountability in Algorithmic Copyright Enforcement*, 19 STAN. TECH. L. REV. 473, 476 (2016).

111. Makena Kelly, *Joe Biden Wants to Revoke Section 230*, VERGE (Jan. 17, 2020, 10:29 AM), https://www.theverge.com/2020/1/17/21070403/joe-biden-president-election-section-230-communications-decency-act-revoke [https://perma.cc/NS7L-AQHA].

112. In one example of this approach, Congress has already amended §230 to remove immunity from claims alleging violations of sex trafficking laws. *See* Allow States and Victims to Fight Online Sex Trafficking Act of 2017, Pub. L. No. 115-164, § 4(a), 132 Stat. 1253, 1254 (2018) (codified at 47 U.S.C. § 230(e)(5) (2018)).

113. Mark Sullivan, *Maybe It's Time to Strip Section 230's Protections for Big Tech*, FAST COMPANY (Nov. 19, 2018), https://www.fastcompany.com/90273352/maybe-its-time-to-take-away-the-outdated-loophole-that-big-tech-exploits [https://perma.cc/WG38-8UCN].

114. Chesney & Citron, *supra* note 31, at 1799.

115. *See* N.Y.U. School of Law, *Legislative Solutions, Individual Rights, and the Question of Government Intervention*, YOUTUBE (June 30, 2020), https://www.youtube.com/watch?v=81Ppe7Vmo8o&list=PLJkLD_s9pYaZU_FpkX_kmH1jh0wjxgM08&index=4 (Gorman noted).

116. Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1603 (2018).

researchers would be allowed access to conduct tests in a confidential setting. Re-thinking of the Computer Abuse Act could also be more constructive.[117] "Infamously problematic,"[118] "the law makes it illegal to access a computer without authorization or in a way that exceeds authorization, but doesn't clearly explain what that means."[119] Amending this law could allow researchers to at least attempt to reverse engineer algorithms to test for flaws without risking liability.

D. Believability

Once makers have created and distributed deepfakes, "audiences, both intended and otherwise, may believe it (or not)."[120] As Matthew Ferraro, Counsel at WilmerHale, puts it, once something gets out, "we can't put the genie back in the bottle."[121] The wave of misinformation that has accompanied COVID-19 underscores the importance of this layer of the framework. So far, there is no concrete evidence that this is the result of coordinated misinformation campaigns—"instead, people are sharing rumors, fake stories, and half-truths about COVID-19 with each other directly . . . as they struggle to understand how best to protect themselves and their families."[122] Our definition of success needs to include not just cutting down the prevalence of deepfakes and other manipulated media but also minimizing our susceptibility to this content.

Why do people believe misinformation? There are many competing and related theories. One main explanation is the effect of "digital echo chambers" and confirmation bias—"the tendency for people to seek and accept information that confirms their existing beliefs while

117. *See* N.Y.U. School of Law, *Legislative Solutions, Individual Rights, and the Question of Government Intervention*, YOUTUBE [hereinafter *Legislative Solutions*] (June 30, 2020), https://www.youtube.com/watch?v=81Ppe7Vmo8o&list=PLJkLD_s9pYaZU_FpkX_kmH1jh0wjxgM08&index=4 (Gorman noted).

118. *Computer Fraud and Abuse Act Reform*, ELECTRONIC FRONTIER FOUND., https://www.eff.org/issues/cfaa (last visited May 13, 2020) [https://perma.cc/2NWU-MEKX].

119. Cindy Cohn & Marcia Hofmann, *Rebooting Computer Crime Law Part 2*, ELECTRONIC FRONTIER FOUND.: DEEPLINKS BLOG (Feb. 4, 2013), https://www.eff.org/deeplinks/2013/02/rebooting-computer-crime-law-part-2-protect-tinkerers-security-researchers [https://perma.cc/MF4P-3UJE].

120. Ayyub, *supra* note 26.

121. *Global Implications*, *supra* note 42.

122. Mark Scott, *Social Media Giants Are Fighting Coronavirus Fake News,* POLITICO (Mar. 12, 2020, 10:31 AM) https://www.politico.com/news/2020/03/12/social-media-giants-are-fighting-coronavirus-fake-news-its-still-spreading-like-wildfire-127038 [https://perma.cc/3NYA-JSW6].

rejecting or ignoring that which contradicts those beliefs."[123] Another is social proof— "when you don't possess sufficient information to solve a given problem, or if you just don't want to or have the time for processing it, then it can be rational to imitate others by way of social proof."[124]

Three interventions aimed at reducing believability were raised by panelists. The first concerns evidence in litigation. In nearly 40 states, lawyers have an ethical duty to stay abreast of changes in law and technology.[125] Competence in identifying and challenging potential deepfakes, or defending authentic evidence from the charge that it is a fake, is an aspect of this competence. Current guidance that litigators should be aware of include model interrogatories that can be used to ask questions about the provenance of digital content as well as tips on litigating authentication rules and recalling digital forensics experts as witnesses.[126] Continued improvements in deepfake detection technology and content verification will help.[127]

Education and media literacy also plays a role. The *Wall Street Journal*, for example, has a media forensics committee made up of editors trained in deepfake detection. This committee hosts regular seminars run by experts, publishes newsroom guidance, and is collaborating with various academic institutions. Similarly, Reuters has collaborated with Facebook's Journalism Project to produce a short course for journalists on manipulated media.[128] Abroad, Sweden's civil-service training manual includes an emphasis on manipulated literacy detection and Ukraine incorporates media-literacy techniques into their public school curriculum.[129] Free online courses like 'Crash

123. *Why We're Susceptible to Fake News, How to Defend Against It*, AM. PSYCHOL. ASS'N (Aug. 10, 2018) https://www.apa.org/news/press/releases/2018/08/fake-news [https://perma.cc/Y8VQ-L5PP].

124. Tom Chatfield, *Why We Believe Fake News*, BBC FUTURE (Sept. 8, 2019), https://www.bbc.com/future/article/20190905-how-our-brains-get-overloaded-by-the-21st-century [https://perma.cc/KS62-JQTW].

125. Robert Ambrogi, *Tech Competence,* LAWSITES, https://www.lawsitesblog.com/tech-competence (last visited Mar. 20, 2020).

126. Lehman et al., *supra* note 27.

127. Blockchain may be part of the solution. *How Blockchains Can and Can't Be Used in Authenticating Video and Countering Deepfakes*, HACKERNOON (Nov. 10, 2018), https://hackernoon.com/how-blockchains-can-and-cant-be-used-in-authenticating-video-and-countering-deepfakes-1c2a59f1f0bd.

128. The full course is available at *Manipulated Media*, REUTERS, https://www.reuters.com/manipulatedmedia (last accessed Mar. 25, 2020).

129. Sasha Ingber, *Students In Ukraine Learn How To Spot Fake Stories, Propaganda And Hate Speech*, NPR (Mar. 22, 2019, 6:25 PM), https://www.npr.org/2019/03/22/705809811/students-in-ukraine-learn-how-to-spot-fake-stories-propaganda-and-hate-speech.

Course Media Literacy' are also becoming more widespread.[130] But one significant barrier to the delivery of these educational programs is how to reach the older population. According to one study, Facebook users the age of 65 and above spread 7 times more 'fake news' than users aged 29 and below.[131]

The final interventions discussed at the believability phase consisted of the technical and design changes to help consumers make more informed decisions and question the believability of information.[132] On March 20th, 2020, Twitter announced that it was working with global public health authorities to identify experts and official health organizations and verify their accounts with the familiar blue checkmark.[133] Other companies are fighting against bot accounts that artificially boost an account's number of followers and likes, in order to disrupt the manufacturing of counterfeit social proof.[134]

E. Impact

The final layer of the disinformation framework is impact: the synthetic media has some impact on society—be it at an individual, an organizational, or global level, which must be captured and measured.[135]

Measuring impact can be difficult. To what extent was disinformation responsible for President Trump's victory in 2016? How many COVID-19 related deaths could have been prevented but-for disinformation? How can we measure loss of trust in institutions, damage to social cohesion, the impact of a wrongful conviction, and the

130. Crash Course Media Studies is available for free. Jay Smooth, CrashCourse, *Media Literacy*, YouTube, https://www.youtube.com/playlist?list=PL8dPuuaLjXtM6jSpzb5gMNsx9kdmqBfmY (last updated Mar. 17, 2020).

131. Andrew Guess et al., *Less than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook*, 5 Sci. Advances 1, 2 (2019), https://advances.sciencemag.org/content/5/1/eaau4586.

132. As Ibrahim noted, a slew of cross-disciplinary experts are conducting research into what kinds of interventions actually work to change people's minds. *See, e.g.*, Antino Kim & Alan R. Dennis, *Says Who? How News Presentation Format Influences Perceived Believability and the Engagement Level of Social Media Users*, 51st Haw. Int'l Conf. on Sys. Sci. 3955, 3962–64 (2018) https://core.ac.uk/download/pdf/143481332.pdf.

133. Twitter Support (@TwitterSupport), Twitter (Mar. 20, 2020, 8:12 PM), https://twitter.com/twittersupport/status/1241155701822476288?lang=en.

134. In California, it is unlawful for any person to use a bot to "communicate or interact with another person in California online," subject to certain provisions, unless the owner of the bot disclosed that this was a bot. Cal. Bus. & Prof. Code §§ 17940–43 (West 2019). For discussion of the law and its contested constitutionality, see Madeline Lamo & Ryan Calo, *Regulating Bot Speech*, UCLA L. Rev. 988, 1008–09, 1017 (2019).

135. Harrison & Aros, *supra* note 33, at 5.

diversion of investment to training and innovation in risk management and fraud protection?[136]

Even if precise measurements escape us, using perceived impact to guide actions can be a helpful safeguard against disproportionate action being taken, and lead to measured responses.[137] For example, the grievous impact of the release of non-consensual deepfake porn on the individual justifies calls for this content to be taken down, by private regulation or force of law, in all circumstances. But there are other situations—for example, when a deepfake intended to critique the government through a satire or parody is at issue—where, even if a piece of content is a proven deepfake, the lesser negative impact, and the importance of a countervailing interest in free speech, might mean that on balance the content should remain available. Deepfake-induced changes to terms of service, as seen at companies like Facebook and Twitter, as well as many of the proposed legislative responses to deepfakes, assume an impact-based balancing approach. Time—and perhaps transparency in decision-making—will tell how these trade-offs are being made.

Conclusion

In the wake of COVID-19, platforms—in particular major information conduits like Twitter, Facebook, Snapchat, Google, and Youtube—have taken an unprecedentedly interventionist approach to content moderation, moving quickly to remove misinformation and adjusting their platform's design to direct users to official sources of information.[138] These moves have earned these companies praise in many quarters,[139] and have prompted some to wonder whether a fundamental shift away from the "techlash" is underway.[140]

But there is another dynamic to consider. This global health crisis has underscored that the impact of misinformation, disinformation,

136. *Id.* at 10–12.

137. N.Y.U. School of Law, *Legislative Solutions, Individual Rights, and the Question of Government Intervention*, YouTube (June 30, 2020), https://www.youtube.com/watch?v=81Ppe7Vmo8o&list=PLJkLD_s9pYaZU_FpkX_kmH1jh0wjxgM08&index=4 (as Volkert underscored).

138. Sarah Kreps & Brendan Nyhan, *Coronavirus Fake News Isn't Like Other Fake News*, Foreign Affairs (Mar. 30, 2020), https://www.foreignaffairs.com/articles/2020-03-30/coronavirus-fake-news-isnt-other-fake-news.

139. Ben Smith, *When Facebook Is More Trustworthy Than the President*, N.Y. Times (Mar. 15, 2020), https://www.nytimes.com/2020/03/15/business/media/coronavirus-facebook-twitter-social-media.html.

140. Steven Levy, *Has the Coronavirus Killed the Techlash,* Wired (Mar. 20, 2020, 9:00 AM), https://www.wired.com/story/plaintext-has-the-coronavirus-killed-the-techlash/.

and malinformation can be profound. What intermediaries choose to do, or not to do, with the content they host can be a matter of life or death. And as social isolation highlights the public's dependence on private intermediaries more visibly than ever before, it may be even more difficult for these companies to disclaim social responsibility and escape scrutiny. Health information, where science can produce an evidentiary standard, is not like political speech, where determining whether or to what extent content is false or misleading can involve difficult subjective judgements.[141] Regardless, having taken such interventionist measures at this point in time, platforms will find themselves having to justify later backpedaling in other areas. The implications for free speech could be immense. The questions raised at this conference—most fundamentally, what content should be allowed, who decides, and how?—are more pressing than ever before.

Stakeholders, ranging from individuals to multi-national corporations and governments need to come together to determine how as a society we want these questions to be answered. "When Seeing Isn't Believing: Deepfakes and the Law" was a worthy contribution to this effort.

141. Kreps & Nyhan, *supra* note 138.