WILL YOU BELIEVE IT WHEN YOU SEE IT? HOW AND WHY THE PRESS SHOULD PREPARE FOR DEEPFAKES

Lauren Renaud*

CITE AS: 4 GEO. L. TECH. REV. 241 (2019)

TABLE OF CONTENTS

I. INTRODUCTION

On April 23, 2013, the Associated Press Twitter account tweeted "Breaking: Two Explosions in the White House and Barack Obama is

* Georgetown Law, J.D. 2019. Lauren Renaud is an attorney with the U.S. Department of Justice. The views expressed in this paper are those of the author and do not necessarily represent the views of the Department of Justice or the United States. The author would like to thank her parents, Laura and Jerry Renaud, for their encouragement and support and Professor Erin Carroll for her insightful feedback on the topic and substance of this paper.

injured."[1] While the claim was false and the result of a hack, its effects were very real: in three minutes the Dow Jones Industrial Average plummeted $136 *billion* in market value.[2] The market recovered after the White House and Associated Press scrambled to refute the claim, but what if the tweet had been equally alarming yet far less rebuttable? A tweet falsely warning of an impending attack on the White House accompanied by a realistic video of terrorists planning for it would lead to a panic that could not be stopped by a single retraction tweet. The technology to fabricate such a video already exists, and future advances will render forgeries undetectable to the human eye.

Government officials and scholars have begun raising the alarm about the danger posed by these forgeries, commonly known as deepfakes, but few have focused on their implications for the press.[3] As deepfakes become more realistic and more widely used, the press will play a unique role due to its ability to give credence to and broadcast a deepfake. The press is also uniquely vulnerable to deepfakes because the news industry relies on the trust of its viewers. Because future deepfakes may affect viewers' perceptions of truth, that trust may be impacted.

The press has traditionally adopted practices to verify source information and media. However, as will be explained, these current practices must be supplemented to adequately address deepfakes. If the press does not increase efforts to prepare for deepfakes, the government will fill the void with measures that are more restrictive and less desirable to the press. Three already proposed government measures would criminalize certain uses of deepfakes and potentially impose criminal liability on journalists who publish them.[4] Additionally, if deepfakes are published as legitimate content, news

---

[1] Max Fisher, *Syrian Hackers Claim AP Hack That Tipped Stock Market By $136 Billion. Is It Terrorism?*, WASH. POST (Apr. 23, 2013), https://www.washingtonpost.com/news/worldviews/wp/2013/04/23/syrian-hackers-claim-ap-hack-that-tipped-stock-market-by-136-billion-is-it-terrorism/?utm_term=.4737b70139d1 [https://perma.cc/SC5Q-EV9Y].

[2] Fisher, *supra* note 1; Shawn Langlois, *This Day In History: Hacked AP Tweet About White House Explosions Triggers Panic* (Apr. 23, 2018, 2:08 PM), https://www.marketwatch.com/story/this-day-in-history-hacked-ap-tweet-about-white-house-explosions-triggers-panic-2018-04-23 [perma.cc/3M68-R4YT].

[3] *See* Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CAL. L. REV. 1753, 1784–85 (2019); Press Release, Senator Mark Rubio, VIDEO: At Intelligence Committee Hearing, Rubio Raises Threat Chinese Telecommunications Firms Pose to U.S. National Security (May 15, 2018), https://www.rubio.senate.gov/public/index.cfm?p=Press-Releases&id=B913F422-DC4F-4F19-A664-D9CE70559F87 [https://perma.cc/9GLA-DQ4S]; *Worldwide Threat Assessment of the US Intelligence Community: Hearing before the S. Select Comm. on Intelligence*, 116th Cong. 66–67 (2019) (statement of Senator Angus King, Member, S. Select Comm. on Intelligence) [hereinafter Worldwide Threat Hearing].

[4] *See infra* Section III.A.1.

organizations will experience a loss of public trust in news—a trust that is fundamental to the news industry's ethos and business model. To avoid such government measures and loss of public trust, the press should begin educating journalists about current deepfake capabilities and experiment with counter-deepfake technology.

II. BACKGROUND

Deepfakes have already begun to impact the press, and their impact will increase as the technology improves and diffuses. This Part will briefly explain deepfake technology, provide an overview of how the technology is being used to impact the press, and illustrate why current digital verification techniques will not be sufficient to address deepfakes.[5] A brief note on terminology: this paper uses "the press" and "news organizations" to collectively refer to news-*producing* entities—*e.g.*, newspapers, TV news, bloggers—and not *conduits* for news, most notably social media platforms.[6]

A. Technical Details

What sets deepfakes apart from previous forgery technology is not just that they are a more realistic product but also how they are created. Current deepfake technology uses artificial intelligence (AI), specifically neural networks, to produce images, audio, and videos (audiovisuals) far more realistically and quickly than a human could create tinkering on Photoshop. Neural networks are complex systems of interconnected processing nodes loosely modeled after the human brain.[7] Neural networks learn to recognize

[5] This paper uses the term "deepfakes" to refer only to artificial intelligence-assisted alteration or generation of images, audio, or video. Therefore, merely duplicating and splicing frames in a video, such as the recent Jim Acosta-White House Intern video, does not constitute a deepfake. *See* Drew Harwell, *White House Shares Doctored Video To Support Punishment Of Journalist Jim Acosta*, WASH. POST (Nov. 8, 2018), https://www.washingtonpost.com/technology/2018/11/08/white-house-shares-doctored-video-support-punishment-journalist-jim-acosta/?utm_term=.f2e2623891f2 [https://perma.cc/8NVG-NTP8].

[6] Though social media companies play an important role in the news cycle, entities serving merely as a conduit for news will be impacted by deepfakes differently than news-producing entities and are thus outside the scope of this paper.

[7] *See* Larry Hardesty, *Explained: Neural Networks*, MIT NEWS (Apr. 14, 2017), http://news.mit.edu/2017/explained-neural-networks-deep-learning-0414 [https://perma.cc/PS2U-G7TM].

patterns from data; generally, the more training data fed into a network, the more accurate the model will be.[8]

The use of neural networks is not a new phenomenon, nor is it exclusive to deepfakes. Facebook's image tagging service, for example, uses neural networks to recognize users' faces in images uploaded to the site.[9] The increasing sophistication of these techniques, however, has allowed networks to be trained beyond merely recognizing patterns to identify content; networks can now work in reverse and use patterns to alter or even generate content, with increasing realism.[10] Though machine learning has not yet achieved true photorealism—a perfect, undetectable fake—the generative adversarial network (GAN) approach is pushing deepfakes towards that goal. The GAN approach involves two neural networks: a generative network which runs the desired model, and a discriminator network that checks the work of the generative network in real-time by assessing the degree to which the generative network succeeded.[11] The discriminator network feeds its results to the generative network which uses the feedback to improve its own output.[12] GANs, in a sense, train themselves and self-perpetuate the arms race between deepfake technology and counter-deepfake technology. The GAN approach has a high potential to improve the accuracy of generated images and eventually generate highly realistic videos.[13]

---

[8] Worldwide Threat Hearing, *supra* note 3, at 82 (statement of Def. Intelligence Agency Dir. Lt. Gen. Robert Ashley) ("How do you get deep fakes that are really, really good? Lots of data—that's how you train your algorithms."); Chesney & Citron, *supra* note 3, at 1759; Will Knight, *Real or Fake? AI Is Making It Very Hard to Know*, MIT TECH. REVIEW (May 1, 2017), https://www.technologyreview.com/s/604270/real-or-fake-ai-is-making-it-very-hard-to-know/ [https://perma.cc/MX6E-QW7Y]; Hardesty, *supra* note 7; Cade Metz & Keith Collins, *How an A.I. 'Cat-and-Mouse Game' Generates Believable Fake Photos*, N.Y. TIMES (Jan. 2, 2018), https://www.nytimes.com/interactive/2018/01/02/technology/ai-generated-photos.html [https://perma.cc/S3TC-WKNW].

[9] Lily Hay Newman, *Facebook Can Even ID You in Photos Where Your Face Isn't Showing*, SLATE (June 23, 2015, 1:54 PM), https://slate.com/technology/2015/06/facebooks-new-machine-vision-algorithm-can-identify-people-without-their-faces.html [https://perma.cc/8EGB-HEDG].

[10] Metz & Collins, *supra* note 8.

[11] Chesney & Citron, *supra* note 3, at 1760; Ian J. Goodfellow, et. al, *Generative Adversarial Networks*, 1–2 (arXiv:1406.2661v1, 2014), https://arxiv.org/pdf/1406.2661.pdf [https://perma.cc/CHX4-H4MG].

[12] Chesney & Citron, *supra* note 3, at 1760; Goodfellow, *supra* note 11 at 1–2.

[13] The GAN approach has already seen limited success in generating videos. *See* Carl Vondrick, et. al., GENERATING VIDEOS WITH SCENE DYNAMICS (2016), http://www.cs.columbia.edu/~vondrick/tinyvideo/ (website created in conjunction for the paper, *Generating Videos with Scene Dynamics*, submitted to the 29th Conference on Neural Information Processing Systems) [https://perma.cc/XW6J-R5XP].

B. Current Capabilities & Potential Uses

As noted, there are no publicly known instances of a perfectly undetectable deepfake, but increasingly convincing audiovisual examples have been produced. Importantly, not all deepfakes are malicious—many are educational or further the arts or other professions.[14] That said, perhaps the most widespread and obscene use of deepfakes is deepfake pornography, popularized in part by a Reddit forum, r/deepfakes, that circulated a tool allowing users to superimpose a celebrity's (or anyone's) face onto pornography videos.[15] Of course, there is now an app for that, several in fact, which use AI to superimpose faces in videos, pornography or otherwise.[16] These browser-based apps make it very easy to produce deepfakes, though the results vary in quality.[17]

If web-based browser apps represent the low end of the sophistication spectrum, university researchers, technology companies, and governments represent the upper end. These entities are likely to have more processing power and better technology, as well as time and access to caches of high-quality photos. Researchers at the University of Washington, for example, created a tool that alters videos to change the speech of the video's speaker.[18] Meanwhile, technology companies have successfully altered audio clips of politicians.[19] Little is publicly known about the United States government's or foreign governments' capabilities to create deepfakes. In 2015, the United States spent roughly $1.1 billion on unclassified artificial intelligence research and development, but it is unknown if any of those funds were spent

[14] Chesney & Citron, *supra* note 3, at 1769–71.

[15] Chesney & Citron, *supra* note 3, at 1763; 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 [https://perma.cc/U7F8-H3WQ].

[16] *See, e.g.*, DeepFakesApp, DEEPFAKESAPP https://deepfakesapp.online (advertising that "DFs may be used to create fake celebrity pornographic videos or revenge porn" on the website's home page) (accessed Oct. 30, 2019) [https://perma.cc/7T5P-9CMQ]; FakeApp, MALAVIDA, https://www.malavida.com/en/soft/fakeapp/#gref (accessed (Oct. 30, 2019) [https://perma.cc/TW2W-QGLU].

[17] A New York Times reporter, for example, was able to create a semi-realistic deepfake video of himself superimposed onto Chris Pratt's body using 1,841 photos of himself. Adam Dodge, et. al., *Using Fake Video Technology to Perpetrate Intimate Partner Abuse*, WITHOUT MY CONSENT 5 (2018), https://withoutmyconsent.org/perch/resources/2018-04-25deepfakedomesticviolenceadvisory.pdf [https://perma.cc/T2ND-NT3W].

[18] Chesney & Citron, *supra* note 3, at 1760.

[19] *Id.* at 1761.

researching deepfake-related technology.[20] However, it is known that the Department of Defense's Defense Advanced Research Project Agency is currently supporting the development of counter-deepfake technology.[21]

There are many potential ways deepfake technology could impact or is impacting journalism. News organizations, for example, have begun to experiment with deepfakes. BuzzFeed's CEO coordinated with director Jordan Peele to create and release what it called a PSA video on deepfakes that altered video and audio of President Barack Obama.[22] Meanwhile, a television news agency in China created an English-speaking "AI-anchor"—a computer-generated news anchor with the "facial expressions and actions of a real person."[23] Although the automated tone of the AI-anchor gives its true nature away, its visuals are impressive and its developers contend the AI-anchor will become more realistic over time.[24]

As the introduction suggested, deepfakes could be used to create fake breaking news which may be unwittingly or intentionally broadcasted.[25] Strategically released deepfakes could amplify their effects. For example, a deepfake video of a police altercation might seek to inflame tensions after a real-life police altercation or, as Senator Marco Rubio of Florida has noted, a

---

[20] GREG ALLEN & TANIEL CHAN, BELFER CTR., ARTIFICIAL INTELLIGENCE AND NATIONAL SECURITY 52 (2017), https://www.belfercenter.org/sites/default/files/files/publication/AI%20NatSec%20-%20final.pdf [https://perma.cc/UM3Z-VDTV].

[21] Matt Turek, *Media Forensics (MediFor)*, DARPA, https://www.darpa.mil/program/media-forensics [https://perma.cc/KM33-KM8L]; Will Knight, *The Defense Department Has Produced The First Tools For Catching Deepfakes*, MIT TECH. REV. (Aug. 7, 2018), https://www.technologyreview.com/s/611726/the-defense-department-has-produced-the-first-tools-for-catching-deepfakes/ [https://perma.cc/BDT7-PJVD].

[22] David Mack, *This PSA About Fake News from Barack Obama Is Not What It Appears*, BUZZFEED (Apr. 17, 2018, 11:26 AM), https://www.buzzfeednews.com/article/davidmack/obama-fake-news-jordan-peele-psa-video-buzzfeed [https://perma.cc/67FX-8A64]. The video showed President Obama speaking about the dangers of deepfakes then transitioned to a side-by-side view of President Obama and Peele speaking the same words. In reality, President Obama never said any of things heard in the video; prior audio and video of him had been altered to make it realistically seem as if he was speaking those words. *Id.*

[23] Merrit Kennedy, *AI News Anchor Makes Debut In China*, NPR (Nov. 9, 2018, 6:11 PM), https://www.npr.org/2018/11/09/666239216/ai-news-anchor-makes-debut-in-china [https://perma.cc/QF4V-74ZZ].

[24] Taylor Telford, *These News Anchors Are Professional and Efficient. They're Also Not Human*, WASH. POST (Nov. 9, 2018, 11:40 AM), https://www.washingtonpost.com/business/2018/11/09/these-news-anchors-are-professional-efficient-theyre-also-not-human/?utm_term=.8ff2bd548b32 [https://perma.cc/3D4Z-SE6T]. Currently, the AI-anchor is trained with live broadcasting videos and social media and requires "only 10 minutes of data to effectively mimic a person's voice." *Id.*

[25] *See supra* Part I. There is already one known instance of a TV news station broadcasting a doctored video that it represented as true. *See infra,* Section III.B.

deepfake image of a political candidate engaging in compromising behavior could be timed to the days prior to an election.[26] Referring to the United States 2020 presidential election and beyond, then-Director of National Intelligence Daniel Coats stated "[a]dversaries and strategic competitors probably will attempt to use deepfakes or similar machine-learning technologies to create convincing—but false—image, audio, and video files to augment influence campaigns directed against the United States and our allies and partners."[27] The topical relevance of certain deepfakes would shorten the timeframe a news organization has in which to verify the media at issue.

Additionally, deepfake technology can be used to target, undermine, or impersonate journalists. Sadly, there has already been at least one instance of this. As a result of her reporting, Indian investigative journalist Rana Ayyub endured hateful commentary online which escalated when a pornographic video with her face superimposed onto another woman went viral.[28] Despite technical confirmation that the video was a fake, the video spread via social media.[29] Deepfakes could also be used to undermine journalists by placing them in compromising situations, such as accepting a bribe or colluding with a politician before a political debate. Deepfakes can also be used to impersonate journalists. Recently, a fake Twitter account impersonated a "Senior Journalist at Bloomberg" and used what is likely an AI-generated image as its Twitter profile picture.[30]

C. Existing Verification Methods Must Be Supplemented

The press has long verified source-provided information and audiovisuals (hereinafter collectively referred to as user-generated content) to ensure its veracity before publication. Indeed, the Society of Professional Journalists' Code of Ethics (Code) instructs "Journalists should . . . [v]erify information before releasing it."[31] The Society notes that the Code is merely a

---

[26] 164 CONG. REC. S5010 (daily ed. July 17, 2018) (statement of Senator Rubio).

[27] Worldwide Threat Hearing, *supra* note 3, at 17 (statement of Dir. Nat'l Intelligence Daniel Coats).

[28] Rana Ayyub, *In India, Journalists Face Slut-Shaming and Rape Threats*, N.Y. TIMES, (May 22, 2018), https://www.nytimes.com/2018/05/22/opinion/india-journalists-slut-shaming-rape.html [https://perma.cc/R5EK-GK2M]; *see also* Danielle Citron, *Sexual Privacy*, 128 YALE L.J. 1870, 1922 (2019).

[29] Ayyub, *supra* note 28.

[30] Glenn Fleishman, *How to Spot the Realistic Fake People Creeping Into Your Timelines*, FAST CO. (Apr. 30, 2019), https://www.fastcompany.com/90332538/how-to-spot-the-creepy-fake-faces-who-may-be-lurking-in-your-timelines-deepfaces [https://perma.cc/4URY-SWG7]; Sean O'Kane (@sokane1), TWITTER (Mar. 27, 2019, 2:54 PM), https://twitter.com/sokane1/status/1111023838467362816 [https://perma.cc/L6J5-DVLA].

[31] *SPJ Code of Ethics*, SOC'Y OF PROF. JOURNALISTS (Sept. 6, 2014, 4:49 PM), https://www.spj.org/ethicscode.asp [https://perma.cc/K3WS-DAD6].

guide, not "a set of rules" or legally enforceable,[32] but its principles—especially regarding verification—are foundational to the journalism field and critical to its success.[33] The Code's instruction to verify user-generated content are reflected in the ethics guidelines of many news entities.[34]

The Code directs journalists to various resources on verification, including the *Verification Handbook: An Ultimate Guideline on Digital Age Sourcing for Emergency Coverage* (Handbook).[35] The Handbook specifically addresses how to verify images and video, but the principles and techniques advocated focus primarily on contextualizing user-generated content and will not be sufficient to verify future deepfakes. The Handbook, for example, suggests videos be verified via a five-step process which involves identifying the source, investigating the source, identifying or confirming the video's location and date, and ensuring the video depicts what it says it depicts.[36] Smart and well-resourced creators, however, could anticipate steps journalists will take to verify and avoid common pitfalls such as using audiovisuals already on the Internet and giving away unwanted clues in the background. AI-generated content is literally tailor-made and, while it is currently visually or forensically detectable, it is constantly being improved. For this reason, current practices alone will be insufficient.

---

[32] *Id.* The Society of Professional Journalists has nearly 250 chapters and 6,000 total members. *About SPJ*, SOC'Y OF PROF. JOURNALISTS, https://www.spj.org/aboutspj.asp [https://perma.cc/K7FC-FMDC]; *Chapters*, SOC'Y OF PROF. JOURNALISTS, https://www.spj.org/chapters.asp [https://perma.cc/KH3A-4LMG]; *See SPJ Code of Ethics*, *supra* note 31.

[33] *See generally* BILL KOVACH & TOM ROSENSTIEL, ELEMENTS OF JOURNALISM (3rd ed., 2013); *see also infra* Section III.B;

[34] *See, e.g.*, *Guidelines on Integrity*, N.Y. TIMES (Sept. 25, 2008), https://www.nytimes.com/editorial-standards/guidelines-on-integrity.html ("it is imperative that The Times and its staff maintain the highest possible standards to insure that we do nothing that might erode readers' faith and confidence in our news columns . . . Images in our pages, in the paper or on the Web, that purport to depict reality must be genuine in every way.") [https://perma.cc/TJS5-YYF6]; *Journalistic Integrity*, WARNER MEDIA GRP. (Nov. 07, 2016), https://www.warnermediagroup.com/company/corporate-responsibility/telling-the-worlds-stories/journalistic-integrity [https://perma.cc/A64P-VNC3]; *Los Angeles Times Ethics Guidelines*, L.A. TIMES (June 16, 2014, 12:00 AM), https://www.latimes.com/la-times-ethics-guidelines-story.html ("Photographs and graphics must inform, not mislead. Any attempt to confuse readers or misrepresent visual information is prohibited.") [https://perma.cc/SHQ5-K9QC].

[35] *See*, EUROPEAN JOURNALISM CENTRE, VERIFICATION HANDBOOK: AN ULTIMATE GUIDELINE ON DIGITAL AGE SOURCING FOR EMERGENCY COVERAGE (Craig Silverman, ed. 2016), https://verificationhandbook.com/downloads/verification.handbook.pdf [https://perma.cc/XEZ6-U2CH].

[36] *Id.* at 47–53.

III. CONSEQUENCES OF FAILING TO ADDRESS THE THREAT

To date, the use of deepfake technology in ways that impact journalism is largely anecdotal. The technology will improve and diffuse, however, and the news industry should not wait until it does to adapt its practices and seriously address deepfakes. This Part highlights several potential consequences that may befall the news industry should it fail to appreciate the threat that deepfakes present.

A. Government Responses

Deepfakes already have Congress's and the intelligence community's attention. Members of Congress have repeatedly warned of the threat deepfakes present and of the government's need to be a part of the solution.[37] Other members have sought to ensure the intelligence community has the legal authority and funding it needs to address deepfakes.[38] The Director of National Intelligence, FBI Director, Defense Intelligence Agency Director, and National Geospatial Agency Director have all expressed their concern about the threat deepfakes pose.[39] This is all to underscore that the Executive and Legislative branches are starting to think about deepfakes, and their concern and motivation to seek solutions will only grow as deepfake technology

---

[37] Worldwide Threat Hearing, *supra* note 3, at 82 (statement of Senator Ben Sasse, Member, S. Select Comm. on Intelligence) ("The asymmetric exposure we have where the barrier to entry for deep fakes technology is so low now – lots of entities short of nation-state actors are going to be able to produce this material and again destabilize not just American public trust but markets very rapidly.")

[38] 164 CONG. REC. S5010 (daily ed. July 17, 2018) (statement of Sen. Rubio); Press Release, Senator Maggie Hassan, Senator Hassan Presses Counterterrorism Official on Size of ISIS, Urges FBI Director to Crack Down on "Deepfakes" (Oct. 11, 2018), https://www.hassan.senate.gov/news/in-the-news/senator-hassan-presses-counterterrorism-official-on-size-of-isis-urges-fbi-director-to-crack-down-on-deepfakes [https://perma.cc/5274-HLGA]; Letter from Members of Congress Adam Schiff, Stephanie Murphy, & Carlos Cabelo to Daniel Coats, Dir. Of Nat'l Intelligence (Sept. 13, 2018), https://murphy.house.gov/uploadedfiles/2018-09_odni_deep_fakes_letter.pdf [https://perma.cc/VZY7-5M2Z].

[39] Press Release, Senator Maggie Hassan, *supra* note 38; OFF. OF THE DIR. OF NAT'L. INTELLIGENCE., THE AIM INITIATIVE 1 (document undated), https://www.dni.gov/files/ODNI/documents/AIM-Strategy.pdf [https://perma.cc/GV8U-27LV]; WWTH Director Ashley; Worldwide Threat Hearing, *supra* note 3, at 71 (NGA Director Robert Cardillo) ("As [deepfake] technology advances – and it will – I do worry about, as a community that needs to seek the truth and then speak the truth, in a world in which we can't agree on what is true our job becomes much more difficult."); *id.* at 82 (statement of Def. Intelligence Agency Dir. Lt. Gen. Robert Ashley) ("Our challenge is how do you build the algorithm to identify the anomaly? Because every deep fake has a flaw, or at least now they do.").

improves. What follows is an analysis of two suggested solutions which may in turn impact journalism.

1. *Criminal Prohibitions of Deepfakes*

Bills banning some form of deepfakes have been proposed in both the California legislature and the United States Congress. As originally proposed, the California bill would make it a misdemeanor to "willfully distribut[e] a deceptive recording that the person knows, or *reasonably should have known*, is a deceptive recording" that is likely to deceive a viewer or "defame, slander or embarrass the subject of the recording."[40] The statute does not define "reasonably should have known," and it is unclear what the standard of care would be. If a journalist follows the verification steps detailed in Part II yet fails to realize a video is a deepfake and publishes it, is she liable? If one hundred journalists believe a video to be real but readily-available software would flag it as a fake, does their failure to use digital verification mean they reasonably should have known it was a fake? This unknown should trouble journalists.

The Malicious Deep Fake Prohibition Act of 2018, introduced in the Senate by Senator Ben Sasse of Nebraska, would make it unlawful to distribute an audiovisual with "[1] actual knowledge that the audiovisual record is a deepfake; and [2] the intent that the distribution of the audiovisual record would facilitate criminal or tortious conduct."[41] The bill's prohibition is

---

[40] A.B. 602, 2019 Leg., Reg. Sess (Cal. 2019), https://leginfo.legislature.ca.gov/faces/billVersionsCompareClient.xhtml?bill_id=201920200AB602&cversion=20190AB60299INT [https://perma.cc/C7ND-EUYA] (emphasis added). The bill has since been amended and was signed into law on October 3, 2019. The final version does not impose criminal liability; instead, it provides a private right of action for victims of nonconsensual deepfake pornography, except material that has "newsworthy value." Its original form is nevertheless discussed because the original proposal may be used as a model by other states in the future. All subsequent references to the California bill are to its original proposed form. *See id.* California has also passed a bill intended to prevent deepfakes from influencing elections. The bill prohibits persons and entities from distributing a deepfake involving an election candidate within 60 days of an election, unless the deepfake includes a disclaimer identifying the content as manipulated. The bill requires actual malice and the intent to injure a candidate's reputation or deceive a voter. Candidate victims can seek injunctive or equitable relief, as well as monetary damages. A.B. 730, 2019 Leg, Reg. Sess., (Cal. 2019), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730 [https://perma.cc/DD6N-XFR8].

[41] Malicious Deep Fake Prohibition Act of 2018, S. 3085, 115th Cong. § 2(a) (2018), https://www.congress.gov/115/bills/s3805/BILLS-115s3805is.pdf [https://perma.cc/9MTP-76BW]. The bill has since expired, but Senator Sasse reportedly intends to reintroduce it. *The Newest Front in the Deepfakes War: Does Congress Need to Step In?*, COUNTABLE (Jan. 31, 2019),

largely superfluous as using a deepfake to facilitate a crime is already a crime.[42] The Act's prohibition should nevertheless be concerning to journalists because it prohibits the distribution of deepfakes to facilitate any tortious conduct—mirroring broad language in the Computer Fraud and Abuse Act (CFAA).[43]

Regarding the CFAA's broad invocation of tortious acts, Professor Orin Kerr commented that "[f]ederal prosecutors have been very creative in coming up with such crimes and torts."[44] Victoria Baranetsky has noted that corporations have successfully used the CFAA to challenge web scraping by "treating the site's terms of service as a contract and prohibiting the act therein."[45] If platforms and websites were to ban deepfakes—or certain uses of deepfakes, as some have already done[46]—in its terms of service, knowingly posting a news story containing a deepfake on those platforms may violate the Act's prohibition under the terms of service-contract theory. Since the Act does not require malicious intent, educational or parody deepfakes may be swept into the prohibition.[47]

For example, if YouTube were to prohibit deepfakes in its terms of service, BuzzFeed would potentially be criminally liable for having posted its highly-viewed deepfake PSA depicting President Obama alerting the public to the dangers of deepfakes.[48] Or, borrowing on an extreme example used by Professor Kerr, BuzzFeed would potentially be committing a felony in violation of the Act if instead of posting the deepfake on YouTube, it had released the video at a party that had music "so loud that the party is tortious under state law, as it's a private nuisance."[49] Since the loud party's purpose is

---

https://www.countable.us/articles/20740-newest-front-deepfakes-war-does-congress-need-step [https://perma.cc/C4WC-ELMB].

[42] Blackmailing an individual, for example, by threatening to release a deepfake would already violate the criminal prohibition on blackmail. *See* 18 U.S.C. § 873 (2018).

[43] *See* 18 U.S.C. § 1030(c)(2)(B)(ii).

[44] Orin Kerr, *Should Congress Pass A "Deep Fakes" Law?*, VOLOKH CONSPIRACY (Jan. 31, 2019, 6:05 PM), https://reason.com/volokh/2019/01/31/should-congress-pass-a-deep-fakes-law [https://perma.cc/6JJX-KJYW].

[45] D. Victoria Baranetsky, *Data Journalism and the Law*, TOW CENTER FOR DIGITAL JOURNALISM (Sept. 19, 2018), https://www.cjr.org/tow_center_reports/data-journalism-and-the-law.php#newsgathering [https://perma.cc/W423-NECB].

[46] *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 [https://perma.cc/6QUL-AL7S].

[47] The First Amendment may protect these uses and will be discussed below.

[48] *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/EX33-D3WJ].

[49] Kerr, *supra* note 44.

to unveil the deepfake, BuzzFeed's distribution of the deepfake at the party would be facilitating tortious conduct. The likelihood that prosecutors would elect to prosecute BuzzFeed for the party under the Act is slim, as are the odds that YouTube will categorically ban deepfakes on its platform,[50] but these scenarios illustrate the broad and potentially unintended effects criminal prohibitions of deepfakes might have.

Additionally, the DEEPFAKES Accountability Act, introduced in the House of Representatives in June 2019, would require deepfake producers to include easily visible disclaimers or watermarks to identify the content as altered.[51] Any person who knowingly fails to comply with the identification requirements would be subject to a fine and/or up to five years in prison if one of four conditions are met: (1) the deepfake depicts sexual acts and is produced with intent to humiliate; (2) the person intended to cause violence or interfere with an election; (3) the noncompliance occurred in the course of criminal conduct relating to fraud or identity theft; or (4) the person is a foreign power or agent of a foreign power engaging in unlawful activity, to include interfering in an election.[52]

It is important to note that the Malicious Deep Fake Prohibition Act expressly exempts "activity protected by the First Amendment;"[53] however, all three legislative proposals could nevertheless potentially chill speech and conduct. The bills, if enacted, would likely face First Amendment challenges. The Senate bill might withstand scrutiny given that it is tied to already unlawful criminal or tortious conduct and expressly exempts First Amendment activity. It is difficult to imagine any of the proposals being upheld if a deepfake involved any category of speech that receives strict scrutiny review (e.g., political speech).[54] For these reasons, their constitutionality is dubious in many situations in which journalists might find themselves. Even still, journalists without access to significant legal support may be afraid to take actions that might contravene the law.

Finally, although the two proposed laws have not yet garnered wide support, that could change in the event of a serious deepfake incident—especially if it occurred during the 2020 election or otherwise impacted national security. A recent Belfer Center report on AI concluded "[t]he bigger and more visible the impacts of AI become (and we argue the impacts are

---

[50] Under current law, YouTube and other platforms are unlikely to categorically ban deepfakes but this may change if Communications Decency Act Section 230 immunity is amended. *See infra* Section III.A.2.

[51] Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019**,** H.R. 3230, 116th Cong. (2019).

[52] *Id.* § 2.

[53] Malicious Deep Fake Prohibition Act of 2018, S. 3085, 115th Cong. § 2(a) (2018).

[54] *See* Chesney & Citron, *supra* note 3, at 1803–04.

likely to be increasingly large and obvious over time) the more policymakers will feel justified in making extreme departures from existing policy."[55] Put simply, "[r]adical technology change begets radical government policy ideas."[56]

2. *Amend Section 230 Immunity*

Section 230 of the Communications Decency Act shields "interactive computer services" from liability resulting from content hosted or shared on its service.[57] At the time of its passage, the prevailing thought was that the possibility of liability would inhibit the growth of the Internet and would disincentivize companies from taking measures to remove obscene content from their services.[58] Twenty-two years later, as Professors Danielle Citron and Bobby Chesney have explained, "Section 230 has evolved into a kind of super-immunity" with the result that "platforms have no liability-based reason to take down illicit material, and. . .victims have no legal leverage to insist otherwise."[59]

Though Section 230 immunity is primarily thought to protect platforms such as Facebook and Yelp, it can protect news organizations in important ways. News sites, for example, enjoy immunity for comments posted to online articles by third-parties.[60] News organizations also enjoy secondary benefits: platforms facilitate free expression, and amending Section 230 immunity may cause platforms to over-moderate content and bar legitimate speech.[61] Taken to the extreme, the litigation risk might lead to platforms removing news articles containing audiovisuals it fears or mistakenly detects are fabricated and potentially unlawful. For these reasons, it is worth briefly exploring why and how Section 230 immunity might be amended to address deepfakes.[62]

---

[55] ALLEN & CHAN, *supra* note 20, at 49.
[56] *Id.* at 3.
[57] 47 U.S.C. § 230 (2018).
[58] Chesney & Citron, *supra* note 3, at 1796–97.
[59] *Id.* at 1798.
[60] *Republication in The Internet Age*, REPORTERS COMM. FOR FREEDOM OF THE PRESS, https://www.rcfp.org/journals/news-media-and-law-summer-2014/republication-internet-age/ (accessed Oct. 30, 2019) [https://perma.cc/6XK3-EXDE].
[61] *See infra* notes 71–72 and accompanying text.
[62] This paper does not take a position on whether Section 230 should be amended; rather, it assumes the possibility and highlights the potential impact reform might have on journalists. For a more detailed analysis on the merits and mechanics of Section 230 immunity, see Danielle Citron & Benjamin Wittes, *The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity*, 86 FORDHAM L. REV. 2 (2017); Citron, *supra* note 28.

Amending Section 230 immunity could be beneficial in addressing deepfakes because it would restore a legal incentive to remove harmful deepfakes (i.e., ones that facilitate illegal or tortious conduct). Senator Mark Warner has highlighted the necessity of this type of legal response noting that platforms are the "least-cost avoiders" in addressing the harm.[63] Although strong market incentives exist to remove harmful deepfakes, these can be insufficient.[64]

Several amendments to Section 230 have already been proposed. One proposal would amend Section 230 such that platforms would be immune only if "they could show that their response to unlawful uses of their services was reasonable."[65] This "reasonableness" standard is meant to mitigate potential concerns that small providers would be unduly burdened because what is "reasonable" for an emerging platform would be different from what is reasonable for a larger platform like Twitter.[66] Moreover, the standard would account for emerging counter-deepfake technologies. If counter-deepfake technologies improve and become widely available, it would become increasingly unreasonable for companies not to employ them.[67] Another proposal would similarly seek to revoke immunity only from "bad" actors with an exemption that is slightly more friendly to platforms.[68] Under this proposal, a platform would enjoy immunity unless it "knowingly and intentionally [left] up unambiguously unlawful content that clearly creates a serious harm to others."[69] Another proposal, even more platform-friendly, would provide for immunity unless the platform "intentionally solicit[ed] or induce[d] illegality or unlawful content."[70]

If Section 230 was amended in a way that incentivized platforms to act on deepfakes, news organizations that publish defamatory (or otherwise harmful) deepfakes could potentially benefit because the platforms' efforts to counter fakes would likely mitigate the harm caused by the defamatory deepfake. This benefit, however, would be marginal and only help those organizations that publish harmful deepfakes. In contrast, potential effects to

---

[63] Mark R. Warner, Potential Policy Proposals for Regulation of Social Media and Technology Firms (July 30, 2018) , https://www.warner.senate.gov/public/_cache/files/d/3/d32c2f17-cc76-4e11-8aa9-897eb3c90d16/65A7C5D983F899DAAE5AA21F57BAD944.social-media-regulation-proposals.pdf (draft white paper) [https://perma.cc/UK2F-RQPL].

[64] For example, one victim of a deepfake pornography video attempted to get the video removed from a website and became the victim of a sextortion attempt. Citron, *supra* note 28 at 40.

[65] Citron & Wittes, *supra* note 62, at 419.

[66] *Id.*

[67] *See id.*

[68] Citron, *supra* note 28, at 63.

[69] *Id.*

[70] *Id.*

free speech and online discourse would affect all news organizations. Legal liability can lead to overzealous removal efforts because "a platform's easiest and cheapest course is to take accusations at face value" and remove the content.[71] The Fourth Circuit has articulated this concern: "Liability upon notice has a chilling effect on the freedom of Internet speech" because of platforms' "natural incentive simply to remove messages upon notification."[72]

It is possible that criminalization of deepfakes and Section 230 immunity reform could occur simultaneously. This would be the worst-case scenario for news organizations because legal liability for harmful deepfakes would likely lead many platforms to ban at least some forms of deepfakes. As previously discussed, the Malicious Deep Fake Prohibition Act combined with more restrictive terms of service could increase liability for news organizations.[73]

Finally, it is important to note that the government has many options other than the two explored here. The government could restrict access to deepfake-related technology via export controls,[74] mandate the use of digital signatures or AI-watermarks,[75] or invest in education to increase the next generation's ability to analyze the credibility of Internet content.[76] These options, however, are less directly applicable to the press and thus were not explored in this analysis.

B. Loss in Public Trust

Public trust in news media is already on the decline. According to a Gallup/Knight Institute survey, sixty-nine percent of American adults say their

---

[71] DAPHNE KELLER, INTERNET PLATFORMS: OBSERVATIONS ON SPEECH, DANGER, AND MONEY 5 (2018), https://www.hoover.org/sites/default/files/research/docs/keller_webreadypdf_final.pdf [https://perma.cc/N9TN-VS5R]. One study that looked at removal in response to copyright claims found that most smaller and lower-profile platforms removed content even when uncertain about the strength of the copyright claim. *Id.*

[72] Zeran v. AOL, 129 F.3d 327, 333 (4th Cir. 1997).

[73] *See supra* Section III.A.1.

[74] ALLEN & CHAN, *supra* note at 20, at 29–30.

[75] *See infra* Section IV.A.2.

[76] Recent studies have shown that adults and minors are generally unable to visually spot fakes. One researcher explained that "[m]ore often than not, people think that the real images are fake and that things that are fake are real. And their confidence is very high. So people are both ignorant and confident, which is the worst combination." Tiffanie Wen, *The Hidden Signs That Can Reveal a Fake Photo*, BBC (June 30, 2017), http://www.bbc.com/future/story/20170629-the-hidden-signs-that-can-reveal-if-a-photo-is-fake [https://perma.cc/8KS2-CC8M].

trust in news media has declined over the past decade.[77] Perhaps the most important finding relevant to deepfakes is that two-thirds of respondents indicated inaccuracy was a reason for their mistrust.[78] Any loss of trust should be concerning to the news industry because consumer trust is a business necessity.[79] The publication of deepfakes—or even the perception that deepfakes are being published—is likely to exacerbate concerns about the inaccuracy of news. Two characteristics of today's climate will help drive this development.

The first characteristic is the increasing tendency to group news organizations together and reinforce monolithic stereotypes (think "mainstream media," "conservative media," and "fake news.") This stereotyping is particularly concerning in the advent of deepfakes because one news organization's blunder may have wider confidence ramifications. NBC News Political Director Chuck Todd recently articulated this effect: "When I make a mistake, it's going to have an impact on Jake Tapper at CNN, it's going to have an impact on Chris Wallace at Fox . . . any error in the 'mainstream media,' we all pay the price."[80]

A recent incident involving doctored video illustrates this effect. In January 2019, a Seattle TV news station broadcasted a doctored video of President Donald Trump's Oval Office address regarding the United States' southern border.[81] In the video, the coloring was saturated to make the president appear orange, his head was enlarged, and his facial expressions

---

[77] Knight Foundation, *Indicators of News Media Trust*, KNIGHT FOUND. (Sept. 11, 2018), https://www.knightfoundation.org/reports/indicators-of-news-media-trust [https://perma.cc/KDD9-YX34].

[78] The open-ended question was: "Thinking now about some of the news media organizations you DO NOT trust, what are some of the reasons why you DO NOT trust those news organizations?" While "two-thirds mentioned accuracy-related reasons at least once," three-quarters of respondents mentioned bias. Interestingly, "Republicans, Democrats, and independents are about equally likely to bring up inaccuracy as a reason they distrust certain news organizations." *Id.*

[79] *See A New Understanding: What Makes People Trust and Rely on News*, AM. PRESS INST. (Apr. 17, 2016), https://www.americanpressinstitute.org/publications/reports/survey-research/trust-news/single-page/ [https://perma.cc/N7FH-H5CJ]; *Guidelines on Integrity*, N.Y. TIMES (Sept. 25, 2008), https://www.nytimes.com/editorial-standards/guidelines-on-integrity.html [https://perma.cc/9E9Z-HAWQ].

[80] Jake Sheridan, *Chuck Todd on Journalism's Long Road Back to Win Public Trust*, DUKE TODAY (Jan. 15, 2019), https://today.duke.edu/2019/01/chuck-todd-journalisms-long-road-back-win-public-trust [https://perma.cc/R6KK-G4CP].

[81] Kyle Swenson, *A Seattle TV Station Aired Doctored Footage Of Trump's Oval Office Speech. The Employee Has Been Fired*, WASH. POST (Jan. 11, 2019), https://www.washingtonpost.com/nation/2019/01/11/seattle-tv-station-aired-doctored-footage-trumps-oval-office-speech-employee-has-been-fired/?utm_term=.573f89517b1a [https://perma.cc/SAV3-AH9Q].

were altered such that he was sticking out his tongue.[82] Viewers took to social media to speculate on whether the video had been doctored until the station confirmed it had aired a doctored version and that the single editor responsible for doing so had been fired.[83] Despite the station's quick response, many Internet users were outraged and, notably, some of their comments were quick to generalize, blaming "the media" writ large: "they wonder why we all think of the Media and News as fake!";[84] "the news media is out of control!"; "the news media are despicably corrupt."[85]

The second characteristic relates to the spread of falsities on social media. A recent study assessing false news on Twitter between 2006 to 2017 found that "falsehood diffused significantly farther, faster, deeper, and more broadly than truth in all categories of information."[86] The study's findings further underscore the heightened responsibility journalists have in the digital age where, though all news spreads quickly, false news spreads even faster. Given this reality, if a deepfake is published and then retracted, it may be more difficult for news of the correction to reach the original readers. Additionally, attempts to undermine journalists would spread rapidly, as journalist Rana Ayyub unfortunately experienced.[87]

IV. SOLUTIONS FOR THE PRESS TO PURSUE

Unfortunately, there is not a single solution to address deepfakes and this paper does not, and could not, proclaim to have the ultimate preparedness plan. While this Part provides a few technology-based solutions, the most

---

[82] *Id.*

[83] *See id.*; *Was video of President Trump's Tuesday address doctored?*, REDDIT (2019), https://www.reddit.com/r/The_Donald/comments/aejp7a/was_video_of_president_trumps_tuesday_address/ [perma.cc link unavailable].

[84] @RSutter, Twitter (Jan. 9, 2019, 11:55 PM), https://twitter.com/RSutter/status/1083225828651978752 [https://perma.cc/Z7BE-YDY6].

[85] Todd Herman, *Q13 FOX Editor Fired Over Doctored Trump Address Video*, 770 KKTH (Jan. 10, 2019 at 1:11 PM), https://mynorthwest.com/1237906/was-video-of-president-trumps-tuesday-address-doctored/?show=comments#comment-4278103291 [https://perma.cc/SQ8J-TSZ9].

[86] Soroush Vosoughi et al., *The Spread Of True And False News Online*, 359 SCI. 1146 (2018), http://science.sciencemag.org/content/359/6380/1146 [https://perma.cc/65VG-TW5C]; Till Daldrup & Francesco Marconi, *How The Wall Street Journal Is Preparing Its Journalists To Detect Deepfakes*, NIEMAN LAB (Nov. 15, 2018, 8:48 AM), http://www.niemanlab.org/2018/11/how-the-wall-street-journal-is-preparing-its-journalists-to-detect-deepfakes/ ("False stories were 70 percent more likely to be retweeted than the truth and reached 1,500 people six times more quickly than accurate articles.") [https://perma.cc/G494-ENPC].

[87] *See* Ayyub, *supra* note 28.

crucial step journalists can take right now is engaging in dialogue, raising awareness, and collaborating within the industry.

A. Technology-based Solutions

There are three promising technological solutions that could aid verification: digital authentication software, digital signatures, and watermarking. Each of these will be explored in turn.

1. *Digital Authentication*

Digital forensics are becoming an increasingly accessible option journalists can use to verify audiovisuals. Current services offer a variety of different tools that use AI, human verifiers, or both. These tools take the form of web-based tools or on-demand verification where a user uploads a specific audiovisual to verify. AI Foundation's Reality Defender software, for example, will soon release a free Google Chrome extension that will scan media that the user encounters while browsing on Chrome and flag media that is manipulated or generated using AI or other means.[88] Another company, Truepic, offers both on-demand and software solutions.[89] Users can drag and drop an image into the Truepic Insight web panel where results on its authenticity will appear instantly.[90] Alternatively, the API version automates the process allowing thousands of images to be quickly verified.[91] In addition to using high-tech verification techniques to flag AI-generated or manipulated content, Truepic's technology also automates many current contextual verification techniques.[92] For example, it can identify whether an image is being repurposed by flagging whether it is an original image (which a journalist might do via Reverse Google Image search), examine metadata to see if it has been manipulated, and flag location spoofing. Finally, while the

[88] *Defend Reality*, AI FOUND., http://www.aifoundation.com/responsibility (accessed Oct. 30, 2019) [https://perma.cc/6DZF-RGWF].
[89] *Our Technology*, TRUEPIC, https://truepic.com/technology/ (accessed Oct. 30, 2019) [https://perma.cc/PFL9-QEJG].
[90] *Id.*
[91] *Id.*
[92] *Id.*

previous solutions have all been services journalists can personally use, the digital verification process can also be outsourced to other entities.[93]

One challenge to using digital verification is ensuring that flagged media has been manipulated in a significant way.[94] Merely opening a photo in Photoshop can alter metadata in a way that software can detect.[95] For this reason, journalists should not automatically discount an audiovisual flagged by software and remember that the technology is an aid, not a final arbiter.

2. *Digital Signatures*

Another promising technology solution is digital signatures. Digital signatures "enable a party to sign a digital object in a way that proves he or she was the one who signed it."[96] A digitally-signed audiovisual would allow viewers to confirm the audiovisual is authentic, was signed by a particular device/person, and has not been modified.[97] Digital signatures can be embedded on devices or in an app. Cannon and Nikon have both implemented the idea in several versions of their cameras, but no smartphone companies are known to have adopted it yet.[98] Serelay, a UK-based company, has an iOS and Android-compatible app that creates digital signatures at the moment of capture when a photo or video is taken using the app. Content taken via the Serelay app can then later be verified either using an on-demand platform or

---

[93] For example, Storyful, a "social media intelligence and news agency," offers user-generated content verification services and has partnered with news organizations such as the *Wall Street Journal* to verify images. *About*, STORYFUL, https://storyful.com/about/ (accessed Nov. 30, 2019) [https://perma.cc/5Q2E-P6VB]; Eamonn Kennedy & Keelan Byrne, *Introducing A New and Improved Storyful Wire*, STORYFUL (Sept. 12, 2018), https://storyful.com/blog/introducing-a-new-and-improved-storyful-newswire/ [https://perma.cc/R2VR-BTM8]; *see* Daldrup & Marconi, *supra* note 86.

[94] *Defend Reality*, supra note 88.

[95] Martin Harran et al., *A Method For Verifying Integrity & Authenticating Digital Media*, 14 SCI. DIRECT, 145, 150 (2018), https://www.sciencedirect.com/science/article/pii/S2210832717300753 [https://perma.cc/CK8S-ASLT].

[96] Herb Lin, *The Danger of Deep Fakes: Responding to Bobby Chesney and Danielle Citron*, LAWFARE BLOG (Feb. 27, 2018, 7:00 AM), https://www.lawfareblog.com/danger-deep-fakes-responding-bobby-chesney-and-danielle-citron [https://perma.cc/KWL7-9B4D].

[97] *Id.* A digital signature is distinct from metadata because a signature would remain intact unless the audiovisual was altered whereas metadata can change merely from being viewed in certain applications. *See* Harran, *supra* note 95.

[98] *Originality Verification Function | OSK-E3*, CANON, http://web.canon.jp/imaging/osk/osk-e3/verifies/index.html (accessed Oct. 30, 2019) [https://perma.cc/6HU4-U4JJ]; *Image Authentication Software*, NIKON, https://imaging.nikon.com/lineup/software/img_auth/index.htm (accessed Oct. 30, 2019) [https://perma.cc/442C-3F7Y]; Lin, *supra* note 96.

an API service similar to Truepic's. If content is flagged as manipulated, the area of manipulation will be highlighted.[99]

3. *AI Watermarks*

There are many responsible uses for AI-generated audiovisuals (educational use, artistic use, etc.), but, given their increasing realism, they can go undetected which could undermine their intended use. AI watermarks can be used to responsibly mark AI-generated audiovisuals so that viewers are aware the content has been computer-generated. The AI Foundation has taken this approach and "partner[ed] with content creators to establish and use an 'Honest AI watermark' to clearly identify and call out AI-generated text images, audio, and video."[100] While malicious actors are unlikely to use AI watermarks, their increased use would, at a minimum, make it easier for journalists and others to quickly establish marked audiovisuals are AI-generated.

B. Education, Dialogue, and Collaboration

While the tools outlined above show promise in aiding digital verification for journalists and Internet users, they cannot supplant current methods. Instead, the tools should be used in addition to contextual verification techniques to most effectively identify deepfakes. As put by *Wall Street Journal* (WSJ) Chief Technology Officer Rajiv Pant: "The way to combat deepfakes is to augment humans with artificial intelligence tools." The most appropriate AI tool and the optimal level of AI assistance will vary according to the audiovisual at issue, how it is being used, and the person or organization that seeks to use it. For this reason, it is imperative that news organizations and journalists begin to examine how deepfakes might affect their organization or products and make individualized assessments about how to best prepare for them. Too few organizations have begun this process.

The WSJ is an outlier and established the WSJ Media Forensics Committee, an "internal deepfakes task force."[101] The Committee is studying current and emerging deepfake technology, evaluating verification practices, and educating its newsroom about deepfakes via training seminars and newsroom guides.[102] Reuters has also publicly acknowledged that it is

---

[99] For a demonstration of what browsing Twitter would look like while using Serelay, see *Verified Twitter Feed*, SERELAY, http://www.verifiedtwitterfeed.com [https://perma.cc/X7JN-TZKV]. For a demonstration of the on-demand service, see *Demo*, SERELAY, https://www.serelay.com/our-products/media/ [https://perma.cc/9WCB-4Y6J].

[100] *Defend Reality*, supra note 88.

[101] Daldrup & Marconi, *supra* note 86.

[102] *Id.*

preparing for deepfakes.[103] Hazel Baker, head of user-generated content news-gathering for Reuters, commented that "[t]here's not a slew of deepfakes on my desk, but I don't want to wait till there are."[104] Reuters has doubled the number of staff verifying video content (from six to twelve) and even worked with a specialist production company to create a deepfake video to test its user-generated content team.[105] Those who were aware the video was manipulated in some manner identified the inconsistencies, while those who were not aware "noticed something was off in the audio but struggled to define it."[106]

News organizations should follow the lead of the WSJ and Reuters and begin thinking about this issue and exploring solutions. If news organizations are planning internally, they should consider publicly identifying their findings and processes to help others prepare because, if the deepfake threat materializes as this author and many others expect, self-preparedness will only go so far. A weakest-link mentality is needed: If the collective news industry is not prepared, each news entity will be affected by the resulting loss in public confidence to the industry, by government action, or both. One scholar has even suggested that a well-prepared news industry may benefit from the advent of deepfakes. He argues that "[d]ire as the case may be, it could offer a great comeback opportunity for mainstream media. As the public learns that it can no longer trust what it sees online, few intermediaries are better placed to function as trusted validators and assessors of mediated reality than professionally trained journalists with access to advanced forensics tools."[107] Benefiting from deepfakes, however, is a lofty goal for an industry that has only just begun to acknowledge them. The news industry currently has a window of opportunity while deepfake technology is relatively nascent and imperfect, but it will quickly close as the technology improves and its wielders become more creative.

V. CONCLUSION

The government measures previously outlined are not inevitable. They, or more extreme measures, are only likely to be implemented if a legal solution seems necessary. Moreover, though the public's perception of

---

[103] *See* Lucinda Southern, *How Reuters Is Training Reporters To Spot "Deepfakes,"* DIGIDAY (Mar. 26, 2019), https://digiday.com/media/reuters-created-a-deepfake-video-to-train-its-journalists-against-fake-news/ [https://perma.cc/C6EF-83G5].

[104] *Id.*

[105] *Id.*

[106] *Id.*

[107] Nicholas Diakopoulos, *Reporting In A Machine Reality: Deepfakes, Misinformation, and What Journalists Can Do About Them*, TOW CENTER FOR DIGITAL JOURNALISM (May 15, 2018), https://www.cjr.org/tow_center/reporting-machine-reality-deepfakes-diakopoulos-journalism.php [https://perma.cc/SQ45-AYLL].

audiovisuals, including those in the news, is likely to change, the news industry can mitigate this by showing it is thinking about these issues and by being prepared when more and better deepfakes circulate. News organizations should prepare by educating their journalists about current deepfake capabilities and assessing which technology-based solutions could best supplement their digital verification practices. We are not yet at an inflection point, but it will not serve the news industry well to wait until we are.