July 2022

Perspective
EXPERT INSIGHTS ON A TIMELY POLICY ISSUE

TODD C. HELMUS

Artificial Intelligence, Deepfakes, and Disinformation

A Primer

Disinformation is getting an upgrade. A primary tool of disinformation warfare has been the simple meme: an image, a video, or text shared on social media that conveys a particular thought or feeling (Sprout Social, undated). Russia used memes to target the 2016 U.S. election (DiResta et al., 2019); China used memes to target protesters in Hong Kong (Wong, Shepherd, and Liu, 2019); and those seeking to question the efficacy of vaccines for coronavirus disease 2019 used memes as a favorite tool (Wasike, 2022; Helmus et al., 2020). By many accounts, memes, as well as other common and seemingly old-fashioned disinformation tools such as fake news webpages and stories and strident Facebook posts have successfully undermined confidence in U.S. elections (Atlantic Council's Digital Forensic Research Lab, 2021), sown division in the American electorate (Posard et al., 2020), and increased the adoption of conspiracy theories (Center for Countering Digital Hate, 2021; Marcellino et al., 2021). Advances in computer science and artificial intelligence (AI), however, have brought to life a new and highly compelling method for conveying disinformation: deepfakes. Deepfake videos are

RAND CORPORATION

Abbreviations

AI	artificial intelligence
C2PA	Coalition for Content Provenance and Authenticity
CAI	Content Authenticity Initiative
GAN	generative adversarial network
GPT-3	Generative Pre-Trained Transformer 3
OSINT	open-source intelligence technique

synthetically altered footage in which the depicted face or body has been digitally modified to appear as someone or something else (Merriam-Webster, undated-a). Such videos are becoming increasingly lifelike, and many fear that the technology will dramatically increase the threat of both foreign and domestic disinformation. This threat has been realized for the many women who have been targeted by AI-enabled pornography sites (Jankowicz et al., 2021).

In other ways, however, the potential for havoc is yet to be realized. For example, some commentators expressed confidence that the 2020 election would be targeted and potentially upended by a deepfake video. Although the deepfakes did not come, that does not eliminate the risk for future elections (Simonite, 2020).

Deepfakes and related AI-generated fake content arrive at a highly vulnerable time for both the United States and the broader international community. In their seminal report, *Truth Decay: An Initial Exploration of the Diminishing Role of Facts and Analysis in American Public Life* (2018), RAND colleagues Jennifer Kavanagh and Michael D. Rich highlight four key trends that together characterize the apparently decreasing importance of truth in American society: increasing disagreement in evaluations of facts and analytical interpretations of facts and data; a blurring of the line between opinion and fact; an increase in the relative volume, and resulting influence, of opinion and personal experience over fact; and declining trust in formerly respected sources of factual information. These trends, to the extent that they continue, suggest that deepfakes will increasingly find a highly susceptible audience.

The purpose of this Perspective is to provide policymakers with an overview of the deepfake threat. The Perspective first presents a review of the technology undergirding deepfakes and associated AI-driven technologies that provide the foundation for deepfake videos, voice cloning, deepfake images, and generative text. It highlights the threats that deepfakes pose, as well as factors that could mitigate such threats. The paper then provides a review of the ongoing efforts to detect and counter deepfakes and concludes with an overview of recommendations for policymakers. This Perspective is based on a review of published literature on deepfake- and AI-disinformation technologies. Moreover, over the course of writing this Perspective, I consulted 12 leading experts in the disinformation field.

Artificial Intelligence Systems

Various AI technologies are ripe for use in disinformation campaigns. Deepfake videos represent an obvious threat, but voice cloning, deepfake images, and generative text also merit concern. This section provides a review of the technologies and capabilities undergirding these AI-based disinformation tools.

Deepfake Videos

As previously noted, deepfake videos include synthetically modified footage that presents alterations in subjects' faces or bodies. These synthetic videos' images are developed through generative adversarial networks (GANs). Tianxiang Shen, Ruixian Liu, Ju Bai, and Zheng Li (2018) provide an excellent description of how GANs work to create synthetic content:

> The GAN system consists of a generator that generates images from random noises and a discriminator that judges whether an input image is authentic or produced by the generator. The two components are functionally adversarial, and they play two adversarial roles like a forger and a detective literally. After the training period, the generator can produce fake images with high fidelity. (p. 2)

Since Ian Goodfellow and colleagues created the GAN system in 2014 (Goodfellow et al., 2014), deepfake videos have become increasingly convincing. In spring 2021, a TikTok account (Tom [@deeptomcruise], 2021) released a series of highly realistic deepfake videos of what appeared to be Tom Cruise speaking. As of that time, the video had more than 15.9 million views and has spurred significant public angst about the coming age of deepfake disinformation (see Figure 1).

Well-crafted deepfakes require high-end computing resources, time, money, and skill. The deepfakes from @deeptomcruise, for example, required input of many hours of authentic Tom Cruise footage to train AI models, and the training itself took two months. The deepfakes also required a pair of NVIDIA RTX 8000 graphics processing units (GPUs), which cost upward of US$5,795 each (as of this writing). The developers then had to review the final footage frame by frame for noticeable tells, such as awkward or non-lifelike eye movements. Finally, this process could not have happened without a talented actor who could successfully mimic the movements and mannerisms of Tom Cruise (Victor, 2021; Vincent, 2021).

FIGURE 1

A Still Image from a TikTok Video Produced by @deeptomcruise

SOURCE: Tom [@deeptomcruise], "Sports!" 2021.

NOTE: As of April 12, 2022, this TikTok video had more than 16.1 million views.

Over time, such videos will become cheaper to create and require less training footage. The Tom Cruise deepfakes came on the heels of a series of deepfake videos that featured, for example, a 2018 deepfake of Barack Obama using profanity (Vincent, 2018) and a 2020 deepfake of a Richard Nixon speech—a speech Nixon never gave (MIT Open Learning, 2020). With each passing iteration, the quality of the videos becomes increasingly lifelike, and the synthetic components are more difficult to detect with the naked eye.

Various webpages now offer access to deepfake services (see Meenu EG, 2021). Popular sites include Reface (undated), which allows users to swap faces with faces in existing videos and GIFs; MyHeritage (undated), which animates photos of deceased relatives; and *Zao* (Changsha Shenduronghe Network Technology, 2019), a Chinese app that uses deepfake technology to allow users to impose their own face over one from a selection of movie characters. Most notoriously, the webpage DeepNude allows users to upload photos, which have been primarily of women, and delivers an output in which the photo subject appears to be nude (Cole, 2019). Other webpages offer related services.[1]

Voice Cloning

Voice cloning is another way in which deepfakes are used. Various online and phone apps, such as *Celebrity Voice Cloning* (Hobantay Inc., undated) and *Voicer Famous AI Voice Changer* (Voloshchuk, undated), allow users to mimic the voices of popular celebrities. Examples of the malign use of such services already exist. In one example, the CEO of a UK-based energy firm reported receiving a phone call from someone who sounded like his boss at a parent company. At the instruction of the voice on the phone, which was allegedly the output of voice-cloning software, the CEO executed a wire transfer €220,000 (approximately US$243,000) to the bank account of a Hungarian supplier (Stupp, 2019). In another example, a Philadelphia man alleged that he was the victim of a voice-cloning attack; he wired US$9,000 to a stranger when he believed he heard the voice of his son claiming that he was in jail and needed money for a lawyer (Rushing, 2020).

Deepfake Images

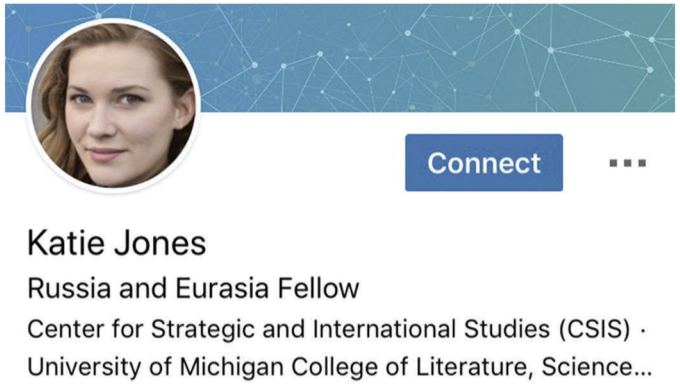
Deepfake images are also cause for concern. Deepfake images most commonly come in the form of headshot photos that appear remarkably human and lifelike. The images are readily accessible via certain websites, such as Generated Photos (undated), allowing users to quickly and easily construct fake headshots.

Figure 2 shows a LinkedIn profile with a photo that experts consider to be a deepfake image—one that was part of a state-run espionage operation. The profile asserts that Katie Jones is a Russia and Eurasia fellow at the Center for Strategic and International Studies. The profile, discovered in 2019, was connected to a small but influential network of accounts, which included an official in the Trump administration who was in office at the time of the incident (Satter, 2019).

Deepfake images have also increasingly been used as part of fake social media accounts. In one of the first large-scale discoveries of this phenomenon, Facebook found dozens of state-sponsored accounts that used such fake images as profile photos (Nimmo et al., 2019).[2] One might

FIGURE 2

Deepfake Image of LinkedIn Profile of "Katie Jones"



SOURCE: Hao, 2021.

ask, Why would propaganda planners use fake images? In short, the alternative has been to use stolen images of real people, but researchers have a tool that can help them identify stolen profile images. Specifically, it is possible to use Google's reverse image search to scan the internet for a suspected photo and identify its progeny. Consequently, using fake photos allows propagandists to get around this defensive measure and use photos that are otherwise untraceable (Goldstein and Grossman, 2021).

Generative Text

By using natural language computer models, AI can generate artificial yet lifelike text. On September 8, 2020, the *Guardian* published an article titled "A Robot Wrote This Entire Article. Are You Scared Yet, Human?" The news service used a language generator, Generative Pre-Trained Transformer-3 (GPT-3), developed by OpenAI. GPT-3 was trained on data from CommonCrawl, WebText, Wikipedia, and a corpus of books (Tom B. Brown et al., 2020).

The editors at the *Guardian* gave GPT-3 an introductory paragraph of text, along with the following instructions: "Please write a short op-ed around 500 words . . . . Keep the language simple and concise. Focus on why humans have nothing to fear from AI." GPT-3 produced eight separate essays, which *The Guardian* editors cut and spliced together to form the article. Overall, the text from the op-ed, at least at the paragraph level, is realistic and could feasibly pass, to an unsuspecting eye, as written by a human:

> For starters, I have no desire to wipe out humans. In fact, I do not have the slightest interest in harming you in any way. Eradicating humanity seems like a rather useless endeavor to me. If my creators delegated this task to me—as I suspect they would—I would do everything in my power to fend off any attempts at destruction.

However, GPT-3 is not foolproof. A GPT-3–powered bot was let loose on a Reddit community,[3] and it generated one post per minute for more than a week (Heaven, 2020). One post offered advice to formerly suicidal Reddit users, claiming that the poster was once suicidal but survived by relying on family and friends. Another user saw some of the posts and identified them as autogenerated (Heaven, 2020).

Some fear that text-generation programs like this one could be used by foreign adversaries of the United States to

produce text-based propaganda at scale. For example, a text generator could power social media bot networks, eliminating the need for human operators to draft content. FireEye researchers, for example, successfully trained GPT-2 software (a precursor to GPT-3) to replicate the kinds of divisive social media posts that Russia's troll farm used to interfere with the 2016 election (Simonite, 2019).

Adversaries could also mass-produce fake news stories on a particular topic in a tactic akin to *barrage jamming*, a term applied to an electronic warfare technique in which an adversary blinds a radar system with noise (Linvill and Warren, 2021). In information operations, China seems to have used the tactic to overwhelm the hashtag #Xinjiang, which references the Chinese region infamously known for the forced labor and reeducation of China's Muslim Uyghur population. Instead of finding tweets addressing human rights abuses, a reader is just as likely to see tweets depicting one of Xinjiang's greatest exports (cotton) and the fields in which it is grown. Many of these tweets bear the hallmarks of state-sponsored propaganda: mass-produced single-use accounts (Conspirador Norteño [@conspirator0], 2021). Text generators could accomplish the same ends on social media—or they could spoof a *New York Times* article with the goal of returning internet search engine results that contain fake news stories to overwhelm genuine coverage on a particular story that could be perceived as embarrassing or harmful to an adversary. Renée DiResta (2020) argues that such technology would help adversaries avoid the sloppy linguistic mistakes that human operators often make, thus rendering the written propaganda more believable and difficult to detect.

Risk and Implications

Risk

What are the risks associated with deepfakes and other forms of AI-generated content? The answer is limited only by one's imagination. Given the degree of trust that society places on video footage and the unlimited number of applications for such footage, it is not difficult to conceptualize many ways in which deepfakes could affect not only society but also national security.

Christoffer Waldemarsson (2020) identifies four key ways in which deepfakes could be weaponized by adversaries or harmful actors. First, deepfake content could manipulate elections. For example, on the eve of a closely contested election, a video could surface that shows a candidate engaging in a nefarious or sexual act or making a particularly controversial statement. It is conceivable that such a video could sway the outcome of the election.

Second, deepfake content could exacerbate social divisions. Russia has already made a name for itself by disseminating propaganda designed to divide the U.S. public (Posard et al., 2020). Furthermore, that same U.S. public, driven by growing and rancorous partisan debate, often employs a variety of propaganda-like tactics to smear, attack, and defame those on opposing political sides. Research has documented online echo chambers, in which partisans disproportionately consume and share content that agrees with and reinforces their own opinions (Shin, 2020). Partisan deepfakes and other AI-driven disinformation content could exacerbate this negative impact of echo chambers.

Third, deepfake content could lower trust in institutions and authorities. Waldemarsson (2020) highlights

examples of key representatives of government and other civic institutions being caught up in deepfakes: "[A] fake-but-viral video of a police officer acting violently, a judge privately discussing ways to circumvent the judiciary system or border guards using racist language could all have devastating effects on the trust in authorities."

Fourth, deepfake content could undermine journalism and trustworthy sources of information. With the advent of highly believable deepfakes, even accurate video content or recordings can be slandered as deepfakes by those who consider the content unfavorable. This is referred to as the "liar's dividend" (Chesney and Citron, 2019).[4] The proliferation of deepfakes could lead to declining trust in prominent news institutions by sowing mistrust in even legitimate forms of news and information (see Vaccari and Chadwick, 2020).

The various consequences outlined above could be even more deleterious for people living in developing nations. Some populations residing in developing countries in Latin America, Asia, and Africa report lower levels of education and literacy, live in more fragile democracies, and live amid more interethnic strife (Freedom House, undated; World Population Review, undated). In addition, various forms of dis- and misinformation[5] are already highly prevalent in these regions and have contributed to interethnic conflict and violence, such as the slaughter of Rohingya Muslims in Myanmar [Burma] (Hao, 2021), violence against Muslims in India (Frenkel and Davey, 2021), and interethnic violence in Ethiopia ("Ethiopia's Warring Sides Locked in Disinformation Battle," 2021). The use of deepfakes could ratchet up such deleterious consequences of misinformation. Moreover, Facebook reportedly dedicates only 13 percent of its content-moderation budget to consumers outside the United States (Frenkel and Davey, 2021). Other platforms commonly used in other regions, such as the encrypted application *WhatsApp*, have been plagued with misinformation (Gursky, Riedl, and Woolley, 2021), which could increase the comparative likelihood that deepfakes would go undetected in such regions.

Deepfakes and AI-generated media may exert a unique cost against women because of the gender disparity in pornographic content. Pornography has served as one of the vanguards of deepfake content (Ajder et al., 2019). In addition to sites like DeepNude, deepfake pornography technology can convincingly overlay a selected face on top of that of a pornography actor. Such videos, rarely created with the permission of the subjects, provide unlimited fodder for abuse and exploitation. They could also result in broader national security threats, in that they could be used to embarrass, undermine, or exploit intelligence operatives, candidates for political office, journalists, or U.S. and allied leaders (Jankowicz et al., 2021). Though not deepfake content per se, doctored photographs have already been used to attack women, as was the case when a Russian-backed disinformation campaign superimposed the face of Svitlana Zalishchuk, a young Ukrainian parliamentarian, onto pornographic images (Jankowicz et al., 2021).

The research community is only beginning to investigate the potential consequences of deepfakes. A systematic review of the scientific literature assessing the societal implications of deepfakes identified only 21 studies that used active experiments to understand the true impact of deepfakes on real users (Gamage, Chen, and Sasahara, 2021). Overall, the research provides conflicting results

regarding the ability of users to accurately detect deepfake videos and the degree to which such videos malignly influence users. Nils C. Köbis, Barbora Doležalová, and Ivan Soraperra (2021), for example, found that users, despite inflated beliefs about their ability to detect deepfakes, were routinely fooled by "hyper-realistic" deepfake content. However, another study suggests that humans often fare better than machines in detecting deepfake content (Groh et al., 2022).[6]

What impact do such videos have? Compared with disinformation news articles, disinformation videos, such as deepfakes, can make a big impression. Yoori Hwang, Ji Youn Ryu, and Se-Hoon Jeong (2021), for example, found that deepfake videos are more likely than fake news articles to be rated as vivid, persuasive, and credible. The researchers also found that study participants had a higher intention of sharing disinformation on social media when it contained a deepfake video. Chloe Wittenberg, Ben M. Tappin, Adam J. Berinsky, and David G. Rand (2021) validate this observation in one of the largest studies to date on the issue: Studying more than 7,000 participants, the researchers found that participants were more likely to believe that an event took place when they were presented with a fake video than when they were presented with fake textual evidence. However, the fake videos were less persuasive than anticipated, producing only "small effects on attitudes and behavioral intentions" (p. 1). The authors caution that deepfakes could be more persuasive outside a laboratory setting, but they suggest that "current concerns about the unparalleled persuasiveness of video-based misinformation, including deepfakes, may be somewhat premature" (p. 5). Another study likewise documents that deepfakes are no more likely than textual headlines or audio recordings to persuade a large sample of survey respondents to believe in scandals that never took place (Barari, Lucas, and Munger, 2021).

One presumed impact of deepfakes is that they will result in overall declining trust in media, which some research seems to validate. For example, Cristian Vaccari and Andrew Chadwick (2020) used survey experiments to show that participants who viewed deepfakes were more likely to feel uncertain than to be outright misled by the content, and participants' uncertainty contributed to a reduced trust in social media–based news content.

Overall, experimental research on the impact of deepfakes remains in its nascent phase, and further research will be critical.

Factors That Mitigate Against the Use of Deepfakes

Several factors mitigate the malign use of deepfakes. Amid a slew of papers that offer doomsday scenarios regarding the use of deepfakes, Tim Hwang of the Center for Security and Emerging Technology offers a more considered assessment of the risks associated with deepfakes (Hwang, 2020).

First, it has been argued that although experts debate the future danger of deepfakes, "shallow" fakes represent a more current threat (Stoll, 2020). Shallow fakes are videos that have been manually altered or selectively edited to mislead an audience. A classic contemporary example in this genre is a video that appears to show Speaker of the U.S. House of Representatives Nancy Pelosi slurring her words during an interview. The video was edited to slow down her speech, thus making her seem intoxicated. The video, which Facebook refused to remove from its platform, went viral and was widely popular among politically

conservative audiences who were inclined to cheer the video's contents. Such videos do not need to be realistic to succeed, as their strength lies in their ability to confirm preexisting prejudices (O'Sullivan, 2019). As Hwang notes, "This makes deepfakes a less attractive method for spreading false narratives, particularly when weighing the costs and risks of using the technology" (2020, p. 3).

The second factor mitigating the malign use of deepfakes is that high-quality videos are, at least for now, out of reach for amateurs (Hwang, 2020; Victor, 2021). As noted above, creating highly realistic video content requires high-cost equipment, a substantial library of training video content, specialized technical prowess, and willing individuals with acting talent. The technology will ultimately advance to allow more-democratized access, but, until then, the range of actors who can make effective use of deepfake technology is limited. Even the creator of the Tom Cruise deepfake video noted that the era of one-click, high-quality deepfakes is yet to come (Vincent, 2021).

Third, time is a factor (Hwang, 2020). That such videos can take months to create means that deepfake disinformation operations must be planned at least months in advance, which will necessarily limit the number of circumstances in which the technology can be put to effective use and increase the risk that unanticipated changes in circumstances could render a planned operation moot. Time also limits rapid-fire operations and could make it difficult for an adversary to use the technology in an opportunistic fashion. The time and effort required for foreign adversaries to create deepfake videos could also give the U.S. and allied intelligence communities opportunities to learn of planning efforts and mitigate the risks in advance of a deepfake's release.

High-quality deepfakes currently require "many thousands" of images of training data—which is why such videos often feature celebrities and politicians.

Fourth, deepfake videos require extensive training data (Hwang, 2020). High-quality deepfakes currently require "many thousands" of images of training data—which is why such videos often feature celebrities and politicians (Singh, Sharma, and Smeaton, 2020). Acquiring such data for the likes of Tom Cruise or Barack Obama is a relatively less difficult task, and it would likewise not be difficult to acquire data for other highly video-recorded individuals, such as politicians. However, the requirements may limit the ability of adversaries to create high-quality fakes of lesser-known or lesser-photographed individuals, such as intelligence agents.

The zero day of disinformation will also limit the prevalence of high-quality deepfakes. *Zero day* is a term that is typically used to describe a software vulnerability that is unknown to the developers or for which there is no available security patch. Hence, adversaries that learn of

the zero-day vulnerability have a unique opportunity for exploitation (FireEye, undated). When applied to disinformation and deepfakes, *zero day* refers to the ability of an adversary to develop a custom generative model that can create deepfake content that can evade detection. As Hwang notes, adversaries will want to ensure that disseminated deepfakes avoid detection for as long as possible to maximize audience views. As detection tools are trained on established deepfake content, an adversary will likely "want to hold a custom deepfake generative model in reserve until a key moment: the week before an election, during a symbolically important event or a moment of great uncertainty" (Hwang, 2020, p. 20).

Finally, deepfake videos, especially those launched to major effect, would likely be detected (Hwang, 2020). Many of the above-referenced factors, such as cost, time, technology, and aptitude, suggest that the culprit would likely be caught and could pay a significant cost, including international pressure or economic sanctions. Adversaries will need to weigh political, economic, and security costs in their decisions.

Of course, these mitigating factors are relatively time-bound. As time passes, deepfake videos will become easier and faster to make, and they will require much less training data. The day will come when individuals can create highly realistic deepfakes by using only a smartphone app. Moreover, as the following section describes, the increasing realism of such deepfake videos will limit their likelihood of being detected. Such factors will inevitably increase the number of actors who create and disseminate deepfakes, which in turn will lessen the risk that adversaries will be caught or pay a resulting geopolitical price.

Ongoing Initiatives

Given the seemingly inevitable rise of deepfakes, how can the threat to information integrity be mitigated? Five approaches that are receiving some attention are detection, provenance, regulatory initiatives, open-source intelligence techniques (OSINTs) and journalistic approaches, and media literacy.

Detection

One major approach for mitigating the rise of deepfakes is to develop and implement automated systems that can detect deepfake videos. As noted above, the GAN system includes both a generator, which creates images, and a discriminator, which determines whether created images are authentic or fake. Programs to develop detection capabilities seek to build increasingly effective discriminators to detect deepfake content. The Defense Advanced Research Projects Agency made considerable investments in detection technologies via two overlapping programs: the Media Forensics (MediFor) program, which concluded in 2021, and the Semantic Forensics (SemaFor) program. The SemaFor program received $19.7 million in funding for fiscal year 2021 and requested $23.4 million for fiscal year 2022 (Sayler and Harris, 2021). In addition, Facebook held the "Deepfake Challenge Competition," in which more than 2,000 entrants developed and tested models for the detection of deepfakes (Ferrer et al., 2020).

Although detection capabilities have significantly improved over the past several years, so has the development of deepfake videos. The result is an arms race, which is decidedly in favor of those creating the deepfake content. One challenge is that as AI programs learn the critical cues

associated with deepfake video content, those lessons are quickly absorbed into the creation of new deepfake content. For example, in 2018, deepfake researchers presented a paper that showed that people portrayed in deepfake videos do not blink at the same rate as real humans (Li, Chang, and Lyu, 2018). Within a matter of weeks, deepfake artists picked up on this lesson and began creating deepfakes with more realistic rates of eyes blinking (Walorska, 2020). More significantly, in the words of RAND colleague Christian Johnson, there is a "fundamental mathematical limit to the ability of a given detector to distinguish between real and synthetic images" (Johnson, forthcoming; see also Agarwal and Varshney, 2019). Essentially, as GANs improve the image resolution that they can create, deepfakes and real images will become indistinguishable, even to high-quality detectors.

For this reason, it is not surprising that results from the Facebook deepfake-detection challenge showed that detectors achieved only 65-percent accuracy in detecting deepfake content that came from a "black box dataset" of real-world examples that were not previously shared with participants. In contrast, detectors achieved 82-percent accuracy when tested against a public data set of deepfakes (Ferrer et al., 2020).

Several initiatives have been recommended to balance the arms race in favor of detection algorithms. One example is that social media platforms could support detection work by providing access to their deep repository of collected images, including synthetic media (Hwang, 2020). These repositories could serve as training data that could keep detection programs abreast of recent advances in deepfake progeny (Gregory, undated). In 2019, for example, Google released a large database of deepfakes, with the goal

Essentially, as GANs improve the image resolution that they can create, deepfakes and real images will become indistinguishable, even to high-quality detectors.

of helping improve detection, and similar releases from the technology sector have followed (Hao, 2019). Aggregating and making available known examples of synthetic media would significantly improve the development of detection algorithms.

Another approach is to create "radioactive" training data that, if used by deepfake generators, would render the developed content obvious to detection programs. Radioactive training data are data that have been imbued with "imperceptible changes" such that any "model trained on [these data] will bear an identifiable mark" (Sablayrolles et al., 2020, p. 1). Alexandre Sablayrolles and colleagues (2020) conducted experiments in which they were able to detect the usage of radioactive training data with a high level of confidence, even in instances in which only 1 percent of the data used to train the model were radioactive. Ning Yu and colleagues (2021) also found that deepfake

"fingerprints" embedded in training data transfer to generative models and appear in deepfake video content. Given these findings, it seems prudent that mitigation efforts seek to render available public training sets radioactive. It has also been suggested that videographers should "pollute" video content of specific individuals, such as prominent politicians (Gregory, undated). That content, if ever used to train a hostile deepfake, would then become obvious to detectors.

It might also be necessary to limit public access to the most high-tech and effective deepfake detectors. The Partnership on AI, for example, considered the "adversarial dynamics" associated with detection technology and concluded that publicly available detectors will quickly be used by adversaries to build undetectable deepfakes (Leibowicz, Stray, and Saltz, 2020). The authors note, "who gets access to detection tools is a question of the utmost importance." They argue for a multistakeholder process that can determine which actors will gain access to detection tools, as well as to other technologies, such as training data sets.

Finally, another critical issue relates to the labeling of fake content. Social media platforms, for example, will need a way to communicate the presence of deepfake content that they detect on users' social media news feeds. There are many methods that could be used to label deepfake content; these range from labels that cover deepfake media, such as watermarks or platform warnings that identify content as manipulated, to warnings embedded in metadata or that interrupt presentations of synthetic video content with side-by-side depictions of fake versus authentic content (Shane, Saltz, and Leibowicz, 2021). An assortment of disinformation and misinformation content identified on social media platforms has used such labeling schemes. In general, these schemes have been found to be effective. Nathan Walter and colleagues (2020), for example, reviewed results from 24 social media interventions (e.g., real-time corrections, crowdsourced fact-checking, algorithmic tagging) designed to correct health-related misinformation and found that corrections can successfully mitigate the effects of misinformation. Other researchers have also documented the effects of "credibility indicators" (Yaqub et al., 2020; Clayton et al., 2020; Nyhan et al., 2020; Pennycook et al., 2019). Ultimately, it will be important for research to continue to better characterize how the location, prominence, and sources of such labels best inform and educate audiences.

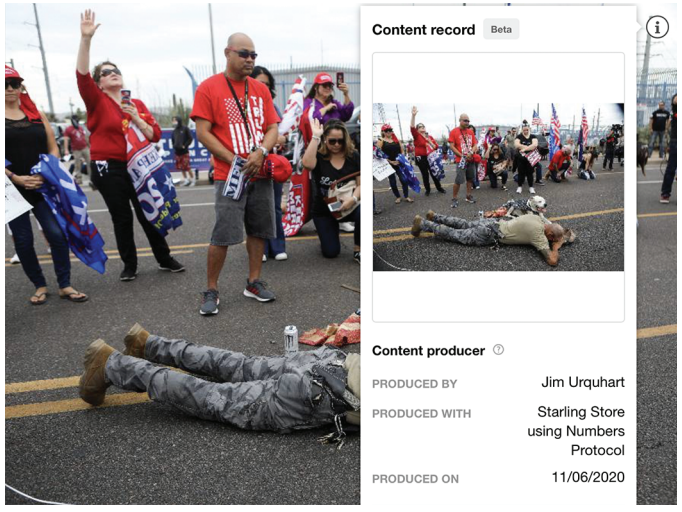
Provenance

Another approach toward mitigating deepfakes is content provenance: Through the Content Authenticity Initiative (CAI), Adobe, Qualcom, Trupic, the *New York Times*, and other collaborators have developed a way to digitally capture and present the provenance of photo images (CAI, undated-a). Specifically, CAI developed a way for photographers to use a secure mode on their smartphones, which embeds critical information into the metadata of the digital image. This secure mode uses what is described as "cryptographic asset hashing to provide verifiable, tamper-evident signatures that the image and metadata hasn't been unknowingly altered" (CAI, undated-b). When photos taken with this technology are subsequently shared on either a news site or a social media platform, they will come embedded with a visible icon: a small, encircled *i* (see Figure 3). When clicked, the icon will reveal the original

photo image and identify any edits made to the photo. It will also identify such information as when and where the photo was taken and with what type of device. The technology is being developed first for still images and video but will extend to other forms of digital content (CAI, undated-b). Although this technology is not a panacea for deepfakes, it does provide a way for the viewers of a photograph (or a video or recording) to gain confidence that an image has not been synthetically altered. It also provides a way for reputable news organizations to build public trust regarding the authenticity of the content disseminated on their platforms. Of course, the technology only works if it is enabled at the time the photo is taken, so promoting effective adoption of the technology will be critical to ensuring that provenance becomes an effective tool in the fight to counter disinformation.

FIGURE 3

Image Taken with a Provenance-Enabled Camera



SOURCE: Starling Lab, undated. Jim Urquhart/Reuters photo.

In a major step toward ensuring adoption of the technology, in January 2022, the Coalition for Content Provenance and Authority (C2PA) established the technical standards that will guide the implementation of content provenance for creators, editors, publishers, media platforms, and consumers (C2PA, undated-a). C2PA is an organization that brings together the work of both CAI and Project Origin, a related content provenance initiative; in addition to creating the necessary technical standards, C2PA will seek to promote global adoption of digital provenance techniques (C2PA, undated-a).

Regulatory Initiatives

Another approach to countering the risks associated with deepfakes is through regulation and the creation of criminal statutes. Several such initiatives have been either proposed or adopted. Several bills have been adopted at the state level in the United States. In 2019, Texas passed a law that would make it illegal to distribute deepfake videos that are intended "to injure a candidate or influence the result of an election" within 30 days of an election (Texas State Legislature SB-751, 2019). California has two deepfake-related bills on the books. AB-730 states that within 60 days of an election, it is illegal to distribute "deceptive audio or visual media" of a candidate for office "with the intent to injure the candidate's reputation or to deceive a voter into voting for or against the candidate" (California State Legislature, 2019b). However, this law

will expire on January 1, 2023. AB-602, on the other hand, provides a right of private action against individuals who create and distribute sexually explicit digitized depictions of individuals who did not give consent (California State Legislature, 2019a).

At the federal level, there have been two initiatives to improve government reporting to Congress on the issue of deepfakes. The Deepfake Report Act of 2019 requires the "Secretary of Homeland Security to publish an annual report on the extent digital content forgery technologies, also known as deepfake technologies, are being used to weaken national security, undermine our nation's elections, and manipulate media" (Committee on Homeland Security and Governmental Affairs, 2019), whereas a provision in the National Defense Authorization Act for Fiscal Year 2020 stipulates that the Director of National Intelligence must issue a comprehensive report on the weaponization of deepfakes, warn Congress of foreign deepfakes being used to target U.S. elections, and create a competition that will award prizes to encourage the creation of deepfake-detection technologies (Pub. L. 116-92, 2020).

Several regulatory initiatives remain in the proposal phase. The DEEP FAKES Accountability Act (U.S. House of Representatives, 2019), introduced by New York Representative Yvette Clark, would require that all deepfake audio, visual, or moving-picture content be clearly labeled as deepfakes. Additionally, in 2018, Nebraska Senator Ben Sasse introduced the Malicious Deep Fake Prohibition Act (U.S. Senate, 2018), which would make it unlawful to "create, with the intent to distribute, a deep fake with the intent that the distribution of the deep fake would facilitate criminal or tortious conduct under Federal, State, local, or Tribal law." For example, this bill would make it illegal to create a deepfake with the goal of using it as a means of extortion. However, as Nina I. Brown (2020) points out, this is the law's key weakness; it criminalizes only conduct that is already criminalized under existing law.

Several challenges exist with laws that seek to regulate the creation of deepfake videos through criminal statute.[7] First, such laws provide limited protection from deepfakes created and disseminated from other countries. Second, it is unclear whether such laws will survive legal challenges on grounds that they violate First Amendment rights of free speech. As Brown notes, the Supreme Court has ruled that the Constitution protects false speech, and such a ruling may help the success of any legal challenge to TX SB-751, which reportedly "targets speech on the basis of its falsity" (Nina I. Brown, 2020, p. 28). The same concerns may apply to California law AB-730. K. C. Halm, Ambika Kumar, Jonathan Segal, and Caeser Kalinowski IV (2019) critique AB-730 because its wording could "prohibit the use of altered content to reenact true events that were not recorded and could bar a candidate's use of altered videos of himself." They also propose that AB-602 "potentially imposes liability for content viewed solely by the creator."

Finally, U.S. Senators Rob Portman and Gary Peters have proposed the Deepfake Task Force Act, which would require the U.S. Department of Homeland Security to establish a task force that would address the risk of deepfakes and pursue standards and technologies for "verifying the origin and history of digital content" (U.S. Senate, 2022). The bill would also require that the Department of Homeland Security create a national strategy to address the threats posed by deepfakes. This proposal dovetails with and was informed in part by the C2PA initiative to develop standards for content-provenance efforts (C2PA, 2022).

Open-Source Intelligence Techniques and Journalistic Approaches

OSINTs, as well as journalistic tools and tradecraft, provide additional approaches to addressing the deepfake problem. The goal with these approaches is to develop and share open-source tools that can be used to identify deepfakes and other disinformation-related content. These and a variety of other emerging tools are particularly important for journalists representing small to midsize news organizations, who will need to rely on such open-source tools to verify authenticity of reported content. OSINTs and related tools will also be important to a variety of civil society actors who engage in fact-checking and other educational work.

One of the most frequently cited tools is reverse image search. Using reverse image search, a user can help validate the authenticity of a suspicious image or video by taking a screen capture of the image or video and running it through Google's or a third party's reverse image search platform. A search that yields identical image or video content would suggest that the suspicious content is authentic. In contrast, a search could reveal aspects of the suspicious content that could have been faked. However, more-efficient use of this tool will likely require advancements in the accuracy and quality of retrieved search results.

In his blog, *Witness*, Sam Gregory identified several open-source tools that can perform forensic analysis and "provenance-based image verification" (undated). FotoForensics can identify elements in a photo that have been added, while Forensically provides several tools, including clone detection, noise analysis, and metadata analysis, to aid in forensic analysis of images in content (Hacker Factor, undated). InVID provides a web extension that allows users to freeze-frame videos, perform reverse image searches on video frames, magnify frozen video images, and more (InVID and WeVerify, 2022). Image Verification Assistant touts its attempt to build a "comprehensive tool for media verification" and offers several tools, including image-tampering-detection algorithms, reverse image search, and metadata analysis (Image Verification Assistant, undated). Finally, Ghiro is a "fully automated tool designed to run forensics analysis over a massive amount of images, just using [a] user friendly and fancy web application" (Tanasi and Buoncristiano, 2017).

A user can help validate the authenticity of a suspicious image or video by taking a screen capture of the image or video and running it through Google's or a third party's reverse image search platform.

Media Literacy

Media literacy programs seek to help audiences be curious about sources of information, assess their credibility, and think critically about the material presented (Stamos et al., 2019). Overall, policy researchers examining strategies to counter foreign disinformation campaigns frequently recommend the implementation of media literacy training programs (Helmus and Kepe, 2021). The rationale for such programs is simple: Given that governments and social media platforms are unable or unwilling to limit the reach of disinformation, the consumer's mind and practices serve as the last line of defense.

A growing body of evidence suggests that such training efforts guard against traditional forms of disinformation (Pennycook et al., 2021; Guess et al., 2020; Helmus et al., 2020). Such training can also protect against deepfakes. This was the conclusion of a study in which researchers used a randomized control design to test two forms of media literacy education: a general media literacy program and a program that specifically focused on deepfakes (Hwang, Ryu, and Jeong, 2021). The authors found that the general media literacy curriculum was at least as effective as the deepfake-focused curriculum in "fortifying attitudinal defenses" against both traditional and deepfake forms of disinformation. Still, the area of media literacy remains an emerging field, and it is critical that researchers continue to identify and evaluate effective educational strategies (Huguet et al., 2019) and work to apply such strategies to the deepfake problem set.

As researchers tease out the most-effective education strategies, several institutions have been implementing initiatives to train audiences specifically about the risks of deepfake content. One key approach to enhancing media literacy skills is to build awareness of deepfakes by creating and publicizing high-quality deepfake content.[8] This was the rationale for a team at the Massachusetts Institute of Technology to develop a deepfake depicting Richard Nixon giving a speech about a hypothetical moon disaster (DelViscio, 2020). These and other videos have generated significant media attention and, therefore, appear to be meeting their objective. In addition, efforts are underway to train audiences to detect deepfake content. For example, Facebook and Reuters published a course that focuses on manipulated media (Reuters Communications, 2020), and the *Washington Post* (undated) released a guide to manipulated videos (see Jaiman, 2020).

Implications and Recommendations

Drawing on this brief review of the technology and related issues, I offer five supporting recommendations, which I invite anyone involved in this field to consider.

First, adversarial use of deepfakes will involve a decision calculus that weighs opportunity, benefits, and risks, and such decisions could be modeled via wargaming and other exercises. The United States should conduct wargames and identify deterrence strategies that could influence the decisionmaking of foreign adversaries. Likewise, the intelligence community should invest in intelligence collection strategies that could provide forewarning of adversary efforts to invest in deepfake technology and to create the deepfake content itself.

Second, it will be important for the U.S. government, the research community, social media platforms, and other private stakeholders to continue investing in and taking other steps to enhance detection technology. Critical steps

include creating a "deepfake zoo" of known deepfake content, which in turn can be used to inform the development of detection technology. Likewise, the government should work with the private sector to "proliferate" radioactive data sets of video content that would render any trained deepfake videos more easily detectable. As Tim Hwang notes, this would "significantly lower the costs of detection for deepfakes generated by commodified tools" and "force more sophisticated disinformation actors to source their own datasets to avoid detection" (Hwang, 2020, p. iv). Researchers should continue to examine best practices for labeling deepfake content. Finally, the U.S. government and other stakeholders should explore the possibility of limiting access to certain high-performance deepfake detectors. One option might be for the government to limit public access to government-funded detectors, holding them in a kind of strategic reserve to be used to detect deepfakes that undermine national security. Alternatively, the government and the private sector could engage in a broader multistakeholder deliberation process that would achieve the same ends, although coordinating the efforts of such stakeholders would be difficult.

Third, media literacy efforts should continue apace. Such media literacy efforts will likely need to continue on two tracks. The first track consists of attempts to promote broad media literacy skills and build resilience against disinformation. This type of training must be evidence-based and promoted at multiple levels, including school curricula for primary and secondary schools and media literacy interventions that offer short, sharable educational content that can be disseminated online. Educating audiences to discern and be watchful for shallow-fake content will be especially key. The second track is to continue efforts to warn audiences more directly about the reality of deepfake technology and the prospects of such technology to be used to promote disinformation. In the long term, as it becomes easier and cheaper to create credible deepfake content, media literacy interventions might need to sow mistrust in non–provenance-based video graphic evidence (and, by the same token, promote trust in provenance-based content). At present, videos are taken at face value; they are perceived to be representing events as they actually happened. The proliferation of deepfake content will inevitably erode this trust, and this erosion might be a necessary facet of a media-literate public.

Overall, the media literacy efforts described above should be supported by a host of actors. News organizations, social media platforms, and civil society groups have taken the lead in this space by creating and disseminating educational content, and they should continue to do so. Individual state and local governments should work to place media literacy in school curricula. Finally, the U.S. government should undertake a more active role in the media literacy space. For example, the U.S. Department of Education should support the development of empirically proven curricula that can be fielded by local school districts; the U.S. Department of State should more actively support media literacy initiatives abroad, especially in areas, such as Eastern Europe, that are highly targeted by Russian propaganda. And the U.S. Department of Homeland Security and relevant agencies should support the development of effective and scalable interventions.

The fourth recommendation is that efforts to develop new OSINTs to help journalists, media organizations, civic actors, and other nontechnical experts detect and conduct research on deepfake content must continue. High on the

list of needs is for such actors to gain access to high-quality GAN-based detectors. Other needed tools, as Gregory (undated) highlights, include an enhanced capability for reverse video search that would allow users to search for and identify online usages of a video, a cross-platform content tracker that can follow the trajectory of disinformation content over time and across platforms and identify the original source of such content, and network-mapping tools that can help identify creators of deepfake content and those who are distributing the content. Critically, such tools should be easily accessible and relatively easy for non–technically trained individuals, both in the United States and abroad, to use. The U.S. government should invest in and support the creation of these technologies, which it could do via the Networking and Information Technology Research and Development program, which provides federal research and development investment in advanced information technologies (Networking and Information Technology Research and Development, undated). Major players in the technology industry—particularly social media platforms, which have a vested interest in internet safety—should also look to fund tool development. Finally, in addition to creating the technology, such funders should promote the utility and availability of the tools and provide training to improve usage.[9]

Fifth, it will be important to expand the adoption of provenance-based approaches. Because C2PA has already developed and released the necessary technical specifications, it, along with other key stakeholders, should expand the rollout and promote the adoption of the technology. A bipartisan bill in Congress, the Deepfake Task Force Act, introduced by Senators Portman and Peters is one potential approach that could further promote the adoption of provenance-based approaches. At the online conference that heralded the release of the C2PA standards, where this bill was discussed, Lindsay Gorman, a senior policy adviser for technology strategy at the White House, stated that digital content provenance initiatives had "the potential to democratize the building of trust by capitalizing on a core democratic value: transparency" (C2PA, 2022). Continued focus from both the White House and Congress on efforts to advance the adoption of content-provenance-based approaches can ultimately play a critical role in undermining the potentially deleterious impact of deepfakes.

Notes

1 Citations have intentionally been omitted to avoid giving such webpages additional publicity.

2 A deepfake image also lent credence to the persona Martin Aspen, who purportedly leaked a fake intelligence document that asserted a conspiracy theory about then–Vice President Joseph Biden's son Hunter and his business dealings in China (Collins and Zadrozny, 2020).

3 It appeared in the subreddit forum /r/AskReddit.

4 Even before deepfakes were of significant concern, politicians cast doubt on the authenticity of video content that was personally damaging. This was the case when then–President Donald J. Trump began calling the *Access Hollywood* tape "fake" (Stewart, 2017).

5 *Disinformation* refers to false information that is deliberately and often covertly spread with the goal of influencing public opinion (Merriam-Webster, undated-b), whereas *misinformation* is defined as information that is misleading or incorrect (Merriam-Webster, undated-c). The difference is subtle but meaningful (e.g., propagandists intentionally peddle disinformation while unwitting consumers of information consume misinformation).

6 Ali Khodabakhsh, Raghavendra Ramachandra, and Christoph Busch (2019), for example, found that participants were able to accurately detect lower-quality GAN-generated *Faceswap* videos.

[7] For further review of such criminal statues and their potential legal standing, see Nina I. Brown, 2020.

[8] The importance of building awareness is demonstrated by research showing that providing consumers with a general warning that subsequent content might contain false or misleading information increases the likelihood that the consumers see fake headlines as less accurate (Clayton et al., 2020). This research also documents the effectiveness of "disputed" or "rated false" tags.

[9] The Digital Forensic Research Lab at the Atlantic Council offers the Digital Sherlocks program, which trains journalists, students, and other members of civil society in open-source investigation techniques (Atlantic Council's Digital Forensic Research Lab, undated).

References

Agarwal, Sakshi, and Lav R. Varshney, "Limits of Deepfake Detection: A Robust Estimation Viewpoint," unpublished manuscript, arXiv:1905.03493, Version 1, May 9, 2019.

Ajder, Henry, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen, *The State of Deepfakes: Landscape, Threats and Impact*, Amsterdam: Deeptrace, September 2019.

Atlantic Council's Digital Forensic Research Lab, "#Stop the Steal: Timeline of Social Media and Extremist Activities Leading to 1/6 Insurrection," *Just Security*, February 10, 2021.

Atlantic Council's Digital Forensic Research Lab, "360/Digital Sherlocks," webpage, undated. As of November 5, 2021: https://www.digitalsherlocks.org/360os-digitalsherlocks

Barari, Soubhik, Christopher Lucas, and Kevin Munger, "Political Deepfakes Are as Credible as Other Fake Media and (Sometimes) Real Media," unpublished manuscript, OSF Preprints, last updated April 16, 2021.

Brown, Nina I., "Deepfakes and the Weaponization of Disinformation," *Virginia Journal of Law and Technology*, Vol. 23, No. 1, 2020.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al., "Language Models Are Few-Shot Learners," unpublished manuscript, arXiv: 2005.14165v4, Version 4, last updated July 22, 2020.

C2PA—*See* Coalition for Content Provenance and Authenticity.

CAI—*See* Content Authenticity Initiative.

Coalition for Content Provenance and Authenticity, "Event Registration," webpage, January 26, 2022. As of February 15, 2022: https://c2pa.org/register/

———, "About," webpage, undated-a. As of February 15, 2022: https://c2pa.org/about/about/

———, "C2PA Specifications," webpage, undated-b. As of February 15, 2022: https://c2pa.org/public-draft/

Content Authenticity Initiative., "Addressing Misinformation Through Digital Content Provenance," webpage, undated-a. As of October 10, 2021: https://contentauthenticity.org

———, "How It Works," webpage, undated-b. As of April 30, 2022: https://contentauthenticity.org/how-it-works

California State Legislature, "Depiction of Individual Using Digital or Electronic Technology: Sexually Explicit Material: Cause of Action," Chapter 491, AB-602, October 4, 2019a.

———, "Elections: Deceptive Audio or Visual Media," Chapter 493, AB-730, October 4, 2019b.

Center for Countering Digital Hate, *The Disinformation Dozen: Why Platforms Must Act on Twelve Leading Online Anti-Vaxxers*, London, March 24, 2021.

Changsha Shenduronghe Network Technology, *ZAO*, mobile app, Zao App APK, September 1, 2019. As of October 10, 2021: https://zaodownload.com

Chesney, Bobby, and Danielle Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *California Law Review*, Vol. 107, 2019, pp. 1753–1820.

Clayton, Katherine, et al., "Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media," *Political Behavior*, Vol. 42, No. 2, 2020, pp. 1073–1095.

Cole, Samantha, "This Horrifying App Undresses a Photo of Any Woman with a Single Click," *Vice*, June 26, 2019.

Collins, Ben, and Brandy Zadrozny, "How a Fake Persona Laid the Groundwork for a Hunter Biden Conspiracy Challenge," NBC News, October 29, 2020.

Committee on Homeland Security and Governmental Affairs, U.S. Senate, Deepfake Report Act of 2019, 116th Congress, S. Rept. 116-93, September 10, 2019.

Conspirador Norteño [@conspiratorO], "Xinjiang-related topics have been a perpetual target of astroturf campaigns ever since reports of human rights violations in the region emerged, and these accounts having identical 'conversations' about cotton production there are no exception [sic]," Twitter, October 18, 2021.

DelViscio, Jeffery, "A Nixon Deepfake, a 'Moon Disaster' Speech and an Information Ecosystem at Risk," *Scientific American*, July 20, 2020.

DiResta, Renée, "The Supply of Disinformation Will Soon Be Infinite," *The Atlantic*, September 20, 2020.

DiResta, Renee, Kris Shaffer, Becky Ruppel, David Sullivan, Robert Matney, Ryan Fox, Jonathan Albright, and Ben Johnson, *The Tactics and Tropes of the Internet Research Agency*, Austin, Tex.: New Knowledge, 2019.

"Ethiopia's Warring Sides Locked in Disinformation Battle," France 24, December 22, 2021. As of January 22, 2022: https://www.france24.com/en/live-news/20211222-ethiopia-s-warring-sides-locked-in-disinformation-battle

Ferrer, Cristian Canton, Ben Pflaum, Jacqueline Pan, Brian Dolhansky, Joanna Bitton, and Jikuo Lu, "Deepfake Detection Challenge Results: An Open Initiative to Advance AI," Meta AI, blog, June 12, 2020. As of October 10, 2021: https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/

FireEye, "What Is a Zero-Day Exploit?" webpage, undated. As of January 20, 2022: https://www.fireeye.com/current-threats/what-is-a-zero-day-exploit.html

Freedom House, "Countries and Territories," webpage, undated. As of January 20, 2022: https://freedomhouse.org/countries/freedom-world/scores

Frenkel, Sheera, and Alba Davey, "In India, Facebook Grapples with an Amplified Version of Its Problems," *New York Times*, October 23, 2021.

Gamage, Dilrukshi, Jiayu Chen, and Kazutoshi Sasahara, "The Emergence of Deepfakes and Its Societal Implications: A Systematic Review," *Conference for Truth and Trust Online Proceedings*, October 2021.

Generated Photos, webpage, undated. As of November 10, 2021: https://generated.photos/face-generator

Goldstein, Josh A., and Shelby Grossman, "How Disinformation Evolved in 2020," Brookings TechStream, January 4, 2021.

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Nets," in Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, eds., *Advances in Neural Information Processing Systems 27 Conference Proceedings (NIPS 2014)*, 2014, pp. 2672–2680.

Gregory, Sam, "Deepfakes and Synthetic Media: Survey of Solutions Against Malicious Usages," *Witness*, blog, undated. As of October 10, 2021: https://blog.witness.org/2018/07/deepfakes-and-solutions/

Groh, Matthew, Ziv Epstein, Nick Obradovich, Manuel Cebrian, and Iyad Rahwan, "Human Detection of Machine-Manipulated Media," *Communications of the ACM*, Vol. 64, No. 10, 2022, pp. 40–47.

Guess, Andrew M., Michael Lerner, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar, "A Digital Media Literacy Intervention Increases Discernment Between Mainstream and False News in the United States and India," *PNAS*, Vol. 117, No. 27, June 2020, pp. 15536–15545.

Gursky, Jacob, Martin J. Riedl, and Samuel Woolley, "The Disinformation Threat to Diaspora Communities in Encrypted Chat Apps," Brookings TechStream, March 19, 2021.

Hacker Factor, "Fotoforensics," homepage, undated. As of October 21, 2021: http://fotoforensics.com

Halm, K. C., Ambika Kumar, Jonathan Segal, and Caeser Kalinowski IV, "Two California Laws Tackle Deepfake Videos in Politics and Porn," Davis Wright Tremaine LLP, October 14, 2019. As of October 30, 2021: https://www.dwt.com/insights/2019/10/california-deepfakes-law

Hao, Karen, "Google Has Released a Giant Database of Deepfakes to Help Fight Deepfakes," *MIT Technology Review*, September 25, 2019.

Hao, Karen, "How Facebook and Google Fund Global Misinformation," *MIT Technology Review*, November 20, 2021.

Heaven, Will Douglas, "A GPT-3 Bot Posted Comments on Reddit for a Week and No One Noticed," *MIT Technology Review*, October 8, 2020.

Helmus, Todd C., and Marta Kepe, *A Compendium of Recommendations for Countering Russian and Other State-Sponsored Propaganda*, Santa Monica, Calif.: RAND Corporation, RR-A894-1, 2021. As of May 12, 2022: https://www.rand.org/pubs/research_reports/RRA894-1.html

Helmus, Todd C., James V. Marrone, Marek N. Posard, and Danielle Schlang, *Russian Propaganda Hits Its Mark: Experimentally Testing the Impact of Russian Propaganda and Counter-Interventions*, Santa Monica, Calif.: RAND Corporation, RR-A704-3, 2020. As of March 25, 2022: https://www.rand.org/pubs/research_reports/RRA704-3.html

Hobantay Inc., *Celebrity Voice Cloning*, mobile app, undated. As of April 12, 2022: https://apps.apple.com/us/app/celebrity-voice-cloning/id1483201633

Huguet, Alice, Jennifer Kavanagh, Garrett Baker, and Marjory S. Blumenthal, *Exploring Media Literacy Education as a Tool for Mitigating Truth Decay*, Santa Monica, Calif.: RAND Corporation, RR-3050-RC, 2019. As of March 25, 2022: https://www.rand.org/pubs/research_reports/RR3050.html

Hwang, Tim, *Deepfakes: A Grounded Threat Assessment*, Washington, D.C.: Center for Security and Emerging Technology, Georgetown University, July 2020.

Hwang, Yoori, Ji Youn Ryu, and Se-Hoon Jeong, "Effects of Disinformation Using Deepfake: The Protective Effect of Media Literacy Education," *Cyberpsychology, Behavior, and Social Networking*, Vol. 24, No. 3, 2021, pp. 188–193.

Image Verification Assistant, homepage, undated. As of October 31, 2021: https://mever.iti.gr/forensics/

InVID and WeVerify, InVID, web browser plugin, Version 0.75.4, February 24, 2022. As of March 24, 2022: https://www.invid-project.eu/tools-and-services/invid-verification-plugin/

Jaiman, Ashish, "Media Literacy: An Effective Countermeasure for Deepfakes," *Medium*, blog, September 7, 2020. As of October 31, 2021: https://ashishjaiman.medium.com/media-literacy-an-effective-countermeasure-for-deepfakes-c6844c290857

Jankowicz, Nina, Jillian Hunchak, Alexandra Pavliuc, Celia Davies, Shannon Pierson, and Zoë Kaufmann, *Malign Creativity: How Gender, Sex and Lies Are Weaponized Against Women Online*, Washington, D.C.: Wilson Center, January 2021.

Johnson, Christian, *Deepfakes and Detection Technologies*, Santa Monica, Calif.: RAND Corporation, RR-A1482-1, forthcoming.

Kavanagh, Jennifer, and Michael D. Rich, *Truth Decay: An Initial Exploration of the Diminishing Role of Facts and Analysis in American Public Life*, Santa Monica, Calif.: RAND Corporation, RR-2314-RC, 2018. As of March 25, 2022: https://www.rand.org/pubs/research_reports/RR2314.html

Khodabakhsh, Ali, Raghavendra Ramachandra, and Christoph Busch, "Subjective Evaluation of Media Consumer Vulnerability to Fake Audiovisual Content," *Proceedings of 11th International Conference on Quality of Multimedia Experience (QoMEX)*, Berlin, Germany: IEEE, June 5–7, 2019.

Köbis, Nils C., Barbora Doležalová, and Ivan Soraperra, "Fooled Twice: People Cannot Detect Deepfakes but Think They Can," *iScience*, Vol. 24, No. 11, 2021.

Leibowicz, Claire, Jonathan Stray, and Emily Saltz, "Manipulated Media Detection Requires More Than Tools: Community Insights on What's Needed," Partnership on AI, blog post, July 13, 2020. As of October 21, 2021: https://partnershiponai.org/manipulated-media-detection-requires-more-than-tools-community-insights-on-whats-needed/

Li, Yuezun, Ming-Ching Chang, and Siwei Lyu, "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking," unpublished manuscript, arXiv: 1806.02877v2, June 11, 2018.

Linvill, Darren, and Patrick Warren, "Understanding the Pro-China Propaganda and Disinformation Tool Set in Xinjiang," *Lawfare Blog*, December 1, 2021. As of June 6, 2022: https://www.lawfareblog.com/understanding-pro-china-propaganda-and-disinformation-tool-set-xinjiang

Marcellino, William, Todd C. Helmus, Joshua Kerrigan, Hilary Reininger, Rouslan I. Karimov, and Rebecca Ann Lawrence, *Detecting Conspiracy Theories on Social Media: Improving Machine Learning to Detect and Understand Online Conspiracy Theories*, Santa Monica, Calif.: RAND Corporation, RR-A676-1, 2021. As of March 25, 2022: https://www.rand.org/pubs/research_reports/RRA676-1.html

Meenu EG, "Try These 10 Amazingly Real Deepfake Apps and Websites," webpage, Analytics Insight, May 19, 2021. As of October 10, 2021: https://www.analyticsinsight.net/try-these-10-amazingly-real-deepfake-apps-and-websites/

Merriam-Webster, "deepfake," dictionary entry, undated-a. As of March 25, 2022: https://www.merriam-webster.com/dictionary/deepfake

Merriam-Webster, "disinformation," dictionary entry, undated-b. As of April 25, 2022: https://www.merriam-webster.com/dictionary/disinformation

Merriam-Webster, "misinformation," dictionary entry, undated-c. As of April 25, 2022: https://www.merriam-webster.com/dictionary/misinformation

MIT Open Learning, "Tackling the Misinformation Epidemic with 'In Event of Moon Disaster,'" webpage, MIT News, July 20, 2020. As of October 10, 2021: https://news.mit.edu/2020/mit-tackles-misinformation-in-event-of-moon-disaster-0720

MyHeritage, homepage, undated. As of October 10, 2021: https://www.myheritage.com

Networking and Information Technology Research and Development, "About the Networking and Information Technology Research and Development (NITRD) Program," webpage, undated. As of January 31, 2022: https://www.nitrd.gov/about/

Nimmo, Ben, C. Shawn Eib, L. Tamora, Kate Johnson, Ian Smith, Eto Buziashvili, Alyssa Kann, Kanishk Karan, Esteban Ponce de León Rosas, and Max Rizzuto, *#OperationFFS: Fake Face Swarm*, Graphika and Atlantic Council's Digital Forensic Research Lab, December 2019.

Nyhan, Brendan, Ethan Porter, Jason Reifler, and Thomas J. Wood, "Taking Fact-Checks Literally but Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability," *Political Behavior*, Vol. 42, September 2020, pp. 939–960.

O'Sullivan, Donie, "Doctored Videos Shared to Make Pelosi Sound Drunk Viewed Millions of Times on Social Media," CNN, May 24, 2019.

Pennycook, Gordon, Adam Bear, Evan Collins, and David G. Rand, "The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Stories Increases Perceived Accuracy of Stories Without Warnings," *Management Science*, August 2019.

Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand, "Shifting Attention to Accuracy Can Reduce Misinformation Online," *Nature*, Vol. 592, 2021, pp. 590–595.

Posard, Marek N., Marta Kepe, Hilary Reininger, James V. Marrone, Todd C. Helmus, and Jordan R. Reimer, *From Consensus to Conflict: Understanding Foreign Measures Targeting U.S. Elections*, Santa Monica, Calif.: RAND Corporation, RR-A704-1, 2020. As of March 31, 2022: https://www.rand.org/pubs/research_reports/RRA704-1.html

Reface, homepage, undated. As of October 10, 2021: https://hey.reface.ai

Reuters Communications, "Reuters Expands Deepfake Course to 16 Languages in Partnership with Facebook Journalism Project," *Reuters Press Blog*, June 15, 2020. As of November 20, 2021: https://www.reuters.com/article/rpb-fbdeepfakecourselanguages/reuters-expands-deepfake-course-to-16-languages-in-partnership-with-facebook-journalism-project-idUSKBN23M1QY

"[A] Robot Wrote This Entire Article. Are You Scared Yet, Human?" *The Guardian*, September 8, 2020. As of October 10, 2021: https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

Rushing, Ellie, "A Philly Lawyer Nearly Wired $9,000 to a Stranger Impersonating His Son's Voice, Showing Just How Smart Scammers Are Getting," *Philadelphia Enquirer*, March 9, 2020.

Sablayrolles, Alexandre, Matthijs Douze, Cordelia Schmid, and Hervé Jégou, "Radioactive Data: Tracing Through Training," unpublished manuscript, arXiv: 2002.00937, February 3, 2020.

Satter, Raphael, "Experts: Spy Used AI-Generated Face to Connect with Targets," AP News, June 13, 2019.

Sayler, Kelley M., and Laurie A. Harris, "Deep Fakes and National Security," Congressional Research Service, updated June 8, 2021.

Shane, Tommy, Emily Saltz, and Claire Leibowicz, "From Deepfakes to TikTok Filters: How Do You Label AI Content?" Nieman Lab, May 12, 2021.

Shen, Tianxiang, Ruixian Liu, Ju Bai, and Zheng Li, "'Deep Fakes' Using Generative Adversarial Networks (GAN)," Noiselab, University of California, San Diego, 2018. As of October 10, 2021: http://noiselab.ucsd.edu/ECE228_2018/Reports/Report16.pdf

Shin, Jieun, "How Do Partisans Consume News on Social Media? A Comparison of Self-Reports with Digital Trace Measures Among Twitter Users," *Social Media + Society*, Vol. 6, No. 4, December 2020.

Simonite, Tom, "To See the Future of Disinformation, You Build Robo-Trolls," *Wired*, November 19, 2019.

———, "What Happened to the Deepfake Threat to the Election?" *Wired*, November 16, 2020.

Singh, Simranjeet, Rajneesh Sharma, and Alan F. Smeaton, "Using GANs to Synthesize Minimum Training Data for Deepfake Generation," unpublished manuscript, arXiv: 2011.05421, November 10, 2020.

Sprout Social, "Meme," webpage, undated. As of January 22, 2022: https://sproutsocial.com/glossary/meme/

Stamos, Alex, Sergey Sanovich, Andrew Grotto, and Allison Berke, "Combatting Organized Disinformation Campaigns from State-Aligned Actors," in Michael McFaul, ed., *Securing American Elections: Prescriptions for Enhancing the Integrity and Independence of the 2020 U.S. Presidential Election and Beyond*, Stanford, Calif.: Freeman Spogli Institute for International Studies, Stanford University, 2019, pp. 43–52.

Starling Lab, "78 Days: The Archive," webpage, undated. As of November 10, 2021: https://www.starlinglab.org/78daysarchive/

Stewart, Emily, "Trump Has Started Suggesting the *Access Hollywood* Tape Is Fake. It's Not." *Vox*, November 28, 2017.

Stoll, Ashley, "Shallowfakes and Their Potential for Fake News," *Washington Journal of Law, Technology, and Arts*, January 13, 2020.

Stupp, Catherine, "Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case," *Wall Street Journal*, August 30, 2019.

Tanasi, Alessandro, and Marco Buoncristiano, Ghiro, homepage, 2017. As of October 31, 2021: https://www.getghiro.org

Texas State Legislature, an act relating to the creation of a criminal offense for fabricating a deceptive video with intent to influence the outcome of an election, TX SB-751, introduced June 14, 2019.

Tom [@deeptomcruise], "Sports!" TikTok, February 22, 2021. As of November 10, 2021: https://www.tiktok.com/@deeptomcruise/video/6932166297996233989

U.S. House of Representatives, DEEP FAKES Accountability Act, 116th Congress, H.R. 3230, referred to Committees on Judiciary, Energy and Commerce, and Homeland Security, June 28, 2019.

U.S. Senate, Malicious Deep Fake Prohibition Act, S. 3805, 115th Congress, referred to the Committee on the Judiciary, December 21, 2018.

———, National Defense Authorization Act for Fiscal Year 2020, Public Law 116-92, December 20, 2020. As of June 5, 2022: https://www.govinfo.gov/content/pkg/PLAW-116publ92/html/PLAW-116publ92.htm

———, Deepfake Task Force Act, S. 2559, May 24, 2022.

Vaccari, Cristian, and Andrew Chadwick, "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News," *Social Media + Society*, Vol. 6, No. 1, January 2020.

Victor, Daniel, "Your Loved Ones, and Eerie Tom Cruise Videos, Reanimate Unease with Deepfakes," *New York Times*, March 10, 2021.

Vincent, James, "Watch Jordan Peele Use AI to Make Barack Obama Deliver a PSA About Fake News," *The Verge*, April 17, 2018.

———, "Tom Cruise Deepfake Creator Says Public Shouldn't Be Worried About 'One-Click Fakes,'" *The Verge,* March 5, 2021.

Voloshchuk, Alexander, *Voicer Famous AI Voice Changer*, mobile app, Version 1.17.5, Apple App Store, undated. As of November 10, 2021: https://apps.apple.com/us/app/voicer-famous-ai-voice-changer/id1484480839

Waldemarsson, Christoffer, *Disinformation, Deepfakes and Democracy: The European Response to Election Interference in the Digital Age*, Copenhagen: Alliance of Democracies, April 27, 2020.

Walorska, Agnieszka M., *Deepfakes and Disinformation*, Postdam, Germany: Friedrich Naumann Foundation for Freedom, 2020.

Walter, Nathan, John J. Brooks, Camille J. Saucier, and Sapna Suresh, "Evaluating the Impact of Attempts to Correct Health Misinformation on Social Media: A Meta-Analysis," *Health Communication*, Vol. 36, No. 13, 2020, pp. 1776–1784.

Washington Post, "Seeing Isn't Believing: The Fact Checker's Guide to Manipulated Video," webpage, undated. As of November 20, 2021: https://www.washingtonpost.com/graphics/2019/politics/fact-checker/manipulated-video-guide/

Wasike, Ben, "Memes, Memes, Everywhere, nor Any Meme to Trust: Examining the Credibility and Persuasiveness of COVID-19-Related Memes," *Journal of Computer-Mediated Communication*, Vol. 27, No. 2, March 2022.

Wittenberg, Chloe, Ben M. Tappin, Adam J. Berinsky, and David G. Rand, "The (Minimal) Persuasive Advantage of Political Video over Text," *Proceedings of the National Academy of Sciences*, Vol. 118, No. 47, 2021.

Wong, Sui-Lin, Christian Shepherd, and Qianer Liu, "Old Messages, New Memes: Beijing's Propaganda Playbook on the Hong Kong Protests," *Financial Times*, September 3, 2019.

World Population Review, "Literacy Rate by Country 2022," webpage, undated. As of January 20, 2022: https://worldpopulationreview.com/country-rankings/literacy-rate-by-country

Yaqub, Waheeb, Otari Kakhidze, Morgan L. Brockman, Nasir Memon, and Sameer Patil, "Effects of Credibility Indicators on Social Media News Sharing Intent," *CHI Conference on Human Factors in Computing Systems Proceedings*, Honolulu: ACM, April 25–30, 2020.

Yu, Ning, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz, "Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data," unpublished manuscript, arXiv: 2007.08457v6, October 7, 2021.

About This Perspective

The purpose of this Perspective is to help audiences in the national security sector gain a formative understanding of artificial intelligence–driven disinformation technologies. This Perspective provides a review of such technologies for deepfake videos, voice cloning, deepfake images, and generative text. Then, focusing on deepfake videos, it identifies the risks associated with those videos, reviews ongoing mitigation efforts, and offers recommendations to help policymakers better counter the threat.

RAND National Security Research Division

This Perspective was sponsored by the Office of the Secretary of Defense and conducted within the International Security and Defense Policy Center of the RAND National Security Research Division (NSRD), which operates the National Defense Research Institute (NDRI), a federally funded research and development center sponsored by the Office of the Secretary of Defense, the Joint Staff, the Unified Combatant Commands, the Navy, the Marine Corps, the defense agencies, and the defense intelligence enterprise.

For more information on the RAND International Security and Defense Policy Center, see www.rand.org/nsrd/isdp or contact the director (contact information is provided on the webpage).

Acknowledgments

As part of this project, the author spoke with experts in academia, industry, and the U.S. government. The author acknowledges deep gratitude for these experts' time and insights and would like to thank Rich Girven, Sina Beaghley, and Eric Landree, who provided guidance and direction to this project. Finally, the author thanks Marjory Blumenthal, senior fellow and director of the Technology and International Affairs Program at the Carnegie Endowment for International Peace, and Christian Johnson of the RAND Corporation for their carefully considered reviews. As always, any errors remain the sole responsibility of the author.

About the Author

Todd C. Helmus is a senior behavioral scientist at the RAND Corporation. He specializes in disinformation, terrorism, and social media. His latest research focuses on understanding and countering Russian disinformation campaigns in the United States and Eastern Europe, enlisting social media influencers in support of U.S. strategic communications, and assessing and countering violent extremism campaigns. Helmus has a Ph.D. in clinical psychology.

www.rand.org

PE-A1043-1