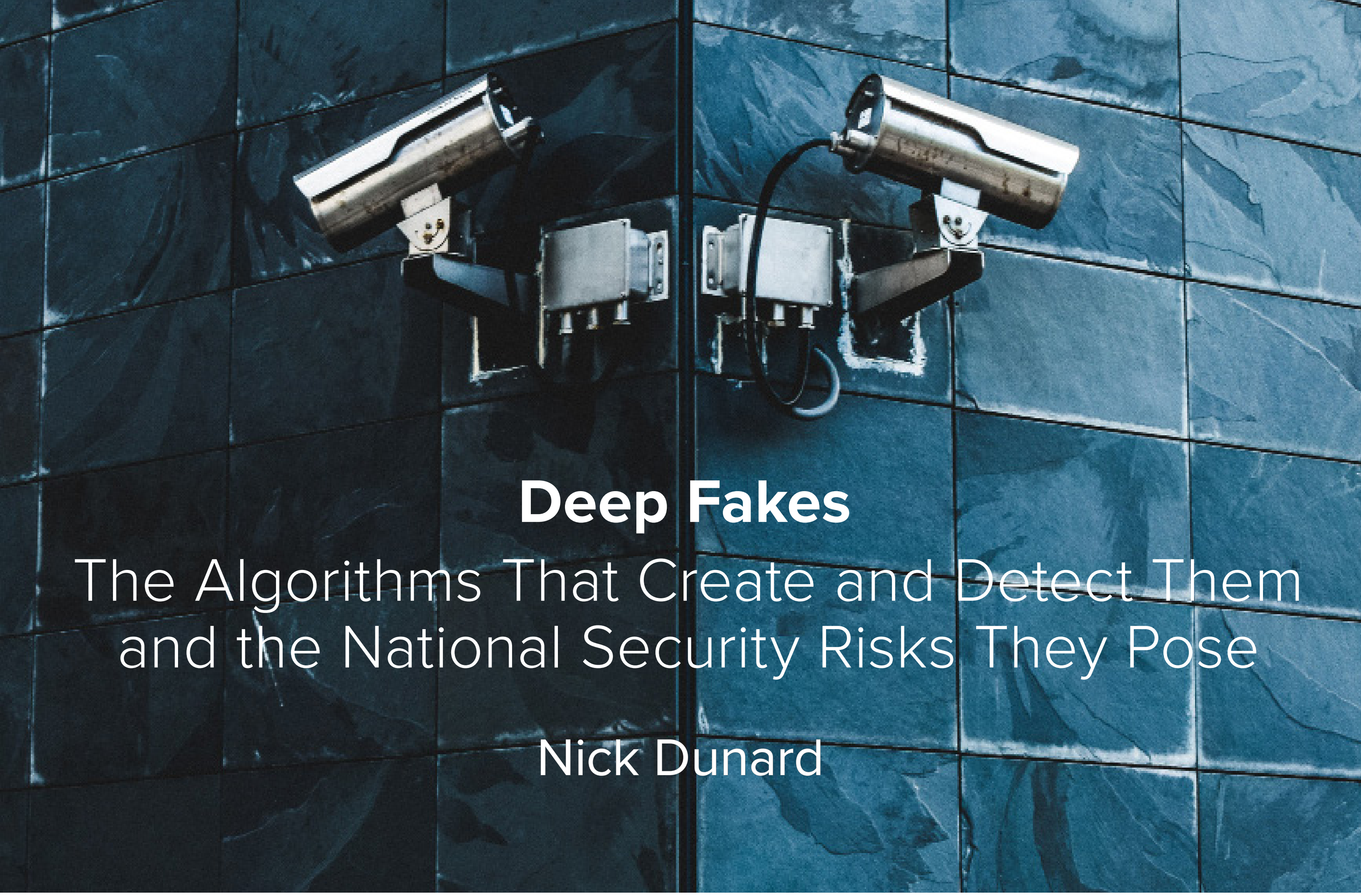
James Madison Undergraduate Research Journal

Volume 8 | Issue 1

2020-2021

Deep Fakes: The Algorithms That Create and Detect Them and the National Security Risks They Pose

Nick Dunard
James Madison University



Deep Fakes

The Algorithms That Create and Detect Them and the National Security Risks They Pose

Nick Dunard

Abstract

The dissemination of deep fakes for nefarious purposes poses significant national security risks to the United States, requiring an urgent development of technologies to detect their use and strategies to mitigate their effects. Deep fakes are images and videos created by or with the assistance of AI algorithms in which a person's likeness, actions, or words have been replaced by someone else's to deceive an audience. Often created with the help of generative adversarial networks, deep fakes can be used to blackmail, harass, exploit, and intimidate individuals and businesses; in large-scale disinformation campaigns, they can incite political tensions around the world and within the U.S. Their broader implication is a deepening challenge to truth in public discourse. The U.S. government, independent researchers, and private companies must collaborate to improve the effectiveness and generalizability of detection methods that can stop the spread of deep fakes.

Keywords: deep fakes, artificial intelligence, machine learning, generative adversarial networks, national security, disinformation, foreign influence operations

The dissemination of deep fakes for nefarious purposes poses significant national security risks to the United States, requiring an urgent development of technologies to detect their use and strategies to mitigate their effects in the public sphere. Deep fakes are images or videos in which a person's likeness, actions, or words have been replaced by someone else's. Deep fakes illustrate how many of the newest national security threats that the United States faces are becoming more technologically advanced and more accessible and easier to operate by motivated individuals and groups.

An interconnected global population and the prevalence of smartphones and social media have enabled unprecedented communication and access to information; these technologies have also opened up new avenues of attack. Known as "disinformation tactics," the threats take many forms, including social media bot networks, fake news stories, blackmail, hacking campaigns, and deep fakes. Disinformation tactics meant to attack an individual or the public's consumption of information and understanding of the world are not new; however, new technologies have allowed their effects to become more widespread and harmful. It is likely that deep fakes will be used in disinformation campaigns by nation state adversaries like Russia and China.

Deep fakes have the potential in both domestic and foreign contexts to sow discord, spread misinformation, damage reputations, and otherwise harm the interests of the United States.

Deep fakes have the potential in both domestic and foreign contexts to sow discord, spread misinformation, damage reputations, and otherwise harm the interests of the United States. Domestically, deep fakes pose a threat to U.S. political and economic processes by targeting specific politicians, business leaders, companies, and news events. They also have a high potential to be used in areas with less technological literacy and areas under less scrutiny by the U.S., where false information can spread for longer periods without being detected. Countering disinformation campaigns and other malicious attacks that utilize deep fakes will require more than just detection methods. Education and communication with the public will need to be part of any counteraction effort. It is in the best interests of the Intelligence Community to understand how deep fakes could be used and how they can be countered.

The artificial intelligence (AI) algorithms and other programs used to create deep fakes become more sophisticated every day, improving their ability to create realistic photos and videos and hampering efforts to create detection algorithms and other countermeasures. Government organizations, independent researchers, and private companies have made significant progress in detecting deep fakes; however, their work has lagged behind the pace with which deep fakes are being developed. There are opportunities to create better, more generalized detection methods to combat the harmful effects of deep fakes.

Deep Fake Creation

Deep fakes are images and videos created by or with the assistance of AI algorithms to deceive an audience. A deep fake could be a realistic photo of a human who does not actually exist, or a video of a public figure saying or doing something they did not actually say or do. The AI algorithms that create deep fakes differ in many ways from the more common applications that AI and machine learning have in our lives. The goal of AI is to create programs able to "learn" in order to solve problems that would ordinarily be too difficult for a computer. AI programs may be presented with data which they learn from to make predictions about previously unseen but related data. For example, a simple AI program may be trained to classify the species of an iris flower. The program does this by learning the features of the data set, in this case the lengths and widths of the petals and sepals, and then using that information to make informed predictions about a new data set. Deep fake creation differs in that the AI does not make a prediction about new data presented to it; instead, it creates new data. These algorithms are known as generative adversarial networks (GANs), and recent innovations have spurred research and development, leading to the emergence of deep fakes

Neural Networks and GANs

The recent history of deep fakes begins with the development of artificial neural networks (ANNs). An ANN is a machine learning model based upon a network of neurons similar to those in the human brain. In the brain, neurons transmit information by producing electrical impulses called action potentials which release neurotransmitters. When a neuron receives enough of these neurotransmitters, it releases its own action potentials, or inhibits itself, instead not firing. These neurons are connected in large networks, allowing complex calculations to be completed. ANNs use this same architecture by connecting networks of artificial neurons which compute huge numbers of possible combinations depending on whether their inputs and outputs are active or not. Today, artificial neural networks are used by businesses and researchers to accomplish highly complex data-intensive tasks such as classifying images, text, and speech.

Building upon ANNs, convolutional neural networks (CNNs) model the ways that the visual cortex processes images to accomplish computer image recognition. Psychologists David H. Hubel and Torsten Wiesel discovered the structure and inner workings of the visual cortex by conducting experiments on cats in 1958 and 1959. Two of their most important discoveries were that many neurons in the visual cortex have small local receptive fields, meaning that they only react to stimuli within a certain region of the visual field, and that some neurons only fire in response to certain orientations of lines and objects while not firing for others. For computer vision and image recognition using CNNs, these discoveries mean that individual neurons in the network do not have to be connected to each other; instead, convolutional layers are used, where neurons only examine and respond to the parts of an image in their receptive fields. This architecture is important for image recognition because of the high number of pixels in any given image. Artificial and convolutional neural networks are the backbone of deep learning programs and inform the algorithms that create deep fakes today: generative adversarial networks.

Artificial and convolutional neural networks are the backbone of deep learning programs and inform the algorithms that create deep fakes today: generative adversarial networks.

Generative adversarial networks were proposed in 2014 by Goodfellow et al. and have ushered in a new era of deep learning research and experimentation, as well as the creation of deep fakes. Goodfellow et al. proposed a new type of generative machine learning model comprised of two competing models: a generative model and a discriminative model. The generative and discriminative models are both neural networks, and if the task is related to image data, it is likely that they are both CNNs. The purpose of the discriminative model is to determine if a sample presented to it, such as an image, is part of an original, real data set, or if it was created by the generative model. The generative model's purpose is to create new samples that could have come from the original, real data set. The models are trained in competition with each other, where the discriminative model attempts to minimize the amount of errors it makes in distinguishing "real" data from "fake" data, and the generative model attempts to maximize that error in the discriminator by creating increasingly better fakes. Since the initial proposal by Goodfellow et al., there has been a surge in research to improve the architecture, efficiency, and realism with which GANs can produce images.

In 2016, Yann Lecun, the VP and Chief AI Scientist at Facebook, described GANs as "the coolest idea in deep learning in the past 20 years."

The idea of placing two neural networks in competition with each other is the most innovative aspect of GANs, as each network improves the other over time. In 2016, Yann Lecun, the VP and Chief AI Scientist at Facebook, described GANs as "the coolest idea in deep learning in the past 20 years." While Goodfellow et al.'s GAN framework was revolutionary, the initial capabilities were quite limited, as the images it generated were very low resolution and often grainy or fuzzy. Three frameworks that have improved the image quality and training stability of GANs since 2014 are deep convolutional GANs, least squares GANs, and StyleGAN. These frameworks are used in the creation of deep fakes, reducing noticeable errors and improving their realism.

Deep Fake Creation with GANs

In practice, if a GAN were trained to generate pictures of cats, the discriminative model would be trained to recognize a cat by using a large data set of different pictures of cats. The generative model would then attempt to create an image that looks like a cat using only random inputs called "noise." At the beginning of the process, the generative model would not be very effective, and the discriminative model would have a high prediction rate between fake and real. As the generative model learns more and more about what the discriminative model looks for to determine if an image is a cat, it can improve its creation of fake cat images. The generative model never actually sees the pictures it creates. Instead, it learns the most important features of the images from information passed by the discriminator tasked with determining whether images are fake or real.

Generative models can be used to produce three main types of deep fake videos: face-swap, lip-sync, and puppet-master. Face-swap videos, which replace a face with another person's face, are the easiest and lowest quality deep fake to produce. Mobile applications such as Snapchat have had similar features for years, and face-swapping is often obvious, as there is usually little done to maintain consistencies such as face movements and position. Lip-sync videos use existing videos of people, and AI manipulates the movements of the mouth to fit new audio. A famous example of a lip-sync deep fake is the Buzzfeed News-produced video of Barack Obama warning of the dangers that deep fakes and disinformation pose, with Jordan Peele serving as the voice actor. The most realistic type of fake videos are puppet-master videos, where a performer acts and says things that they want the target to appear to be doing. Then, using AI tools, the video is used to animate the target as having said and done what the performer did.

National Security Risks

Deep fakes pose national security risks to both individuals and society as a whole in both foreign and domestic contexts. While fake images present risks, fake video and audio allow greater flexibility and therefore pose greater threats. Individuals targeted by deep fakes face reputational harm, loss of employment, and theft and identity fraud. They also may feel threatened and powerless to respond or disprove the fakes. At a society-wide level, deep fakes can be used to spread disinformation; inflame racial, ethnic, cultural, and political tensions; influence election outcomes; and destabilize the U.S. economy. Changing socio-political developments like COVID-19, nationwide racial justice protests, and national elections exacerbate existing political tensions, opening new avenues for disinformation tactics targeting the public.

Deep fakes can be used to spread disinformation; inflame racial, ethnic, cultural, and political tensions; influence election outcomes; and destabilize the U.S. economy.

The American public first truly became aware of online influence operations and disinformation campaigns after the Russian government's "sweeping and systematic" interference in the 2016 U.S. presidentialelection, when Russian operatives hacked and disseminated a candidate's emails and spread fake news through social mediaaccounts. The Intelligence Community has assessed that foreign actors continued their election interference schemes in the 2018 U.S. Congressional Elections and in the 2020 U.S. presidential election. And as the COVID-19 pandemic spread across the globe and within the U.S. in early 2020, intelligence officials watched Chinese operatives orchestrate mass texts to millions of Americans warning of an impending lockdown and martial law, showing the range of options in disinformation campaigns. The U.S. Intelligence Community currently considers efforts like these a top priority, listing them second in the 2019 Worldwide Threat Assessment. Deep fakes may exacerbate theproblems associated with foreign electoral interference, as they provide unprecedented realism to false information.

Deep Fake Photos

Deep fake photos pose significant national security risks for individuals including extortion, identity theft and fraud, and reputational harm. Additionally, they can be used to bolster other elements of a disinformation campaign, such as creating more realistic fake profiles and infiltrating social networks and organizations. These photos are likely to be created by foreign nation states, hacking groups, and aggrieved individuals depending on the purpose, context, and targets.

Individuals who hold positions of power within the U.S. government, private corporations, and large organizations may be blackmailed, extorted, or threatened with deep fake photos. A fake image of someone engaging in drug use or other questionable activities can be used to leverage information, money, or other things of value. According to a Congressional Research Service report, foreign intelligence operatives have already begun using deep fakes in social media profiles to recruit sources in the U.S.

Deep fake photos may also be used to improve the realism of other elements of disinformation and online influence operations. In 2019, researchers discovered a LinkedIn profile of a woman named Katie Jones who appeared to be deeply connected to many national security experts and other political figures in Washington. In fact, no such Katie Jones exists, and many elements of her profile indicate that she was likely created by a GAN. A scaly effect on her ear, mismatched and monochromatic eyes, a blurry earring, and the indistinct background made it clear that the photo was not real. However, the profile was still able to connect with more than 50 users on LinkedIn, including a deputy assistant secretary of state. Similar fake accounts are likely to be used to connect with influential members of government and business to siphon confidential information, create compromising situations, or to recruit them to directly work with foreign governments. While the Katie Jones profile was detected quickly due to the low quality of the image, more sophisticated efforts to fine-tune the generation algorithm can produce fake images that fool the naked eye.

Deep Fake Video and Audio

Deep fake video and audio productions are more likely to pose serious national security risks at a society-wide level than photos due to the limitless possibilities of what can be created and shown. While the most serious threats are likely to be in the domestic context, such as those that target our elections and economy or try to spark hatred and division, serious threats could emerge from the spread of fake videos targeted at individuals or in other countries. The technology used to create these videos has only gotten better, and it may now be used to interfere in deeper, more sinister ways against the U.S. and its citizens.

In more personal or intimate contexts, deep fake videos can be used to harass and intimidate individuals with blackmail or revenge porn. The first deep fake videos emerged in 2017 when internet users interposed the faces of celebrities between those of actors in pornographic videos. Since then, researchers have found that over 90% of deep fake videos are non-consensual porn, mainly targeting women. A report by the Cyber Civil Rights Initiative indicates that 90% of revenge porn victims are women, and that many have suffered reputational and emotional consequences as a result. Recent mobile and computer applications make it easy to create these videos: typically, just a few pictures from social media accounts are enough. Online communities have formed to share and request porn deep fakes of individuals and celebrities, which normalizes the behavior for would-be perpetrators. In one case, a mother of a high school cheerleader created deep fake videos and photos showing her daughter's rivals on the team naked while smoking and drinking to get them kicked off of the team. A concern regarding deep fakes videos being used in local or individual contexts is the difficulty of proving that they are fake. Without the resources that researchers and media organizations can bring to bear, individuals are susceptible to reputational harm, shame, and harassment. While instances like the one above do not pose significant national security threats to the U.S., they do create serious civil liberties and privacy concerns and are likely to be the majority of cases involving deep fakes.

Deep fake videos provide unprecedented customization, targeting, and believability to hostile foreign actors working to spead disinformation.

Deep fake videos can target politicians, business leaders, minority groups, activists, celebrities, members of our armed forces, or anyone else in a position of power or influence. Politicians could be displayed taking bribes or saying racist phrases, and the CEO of a company could be heard talking about a coming recession, triggering panic selling in the market. A video could circulate of police officers indiscriminately assaulting innocent civilians, causing riots across the nation before it can be disproven. According to Special Counsel Robert S. Mueller III in his 2019 *Report on the Investigation into Russian Interference in the 2016 Presidential Election*, a major element of the Russian social media campaign in the months leading up to the election was to "provoke and amplify political and social discord in the United States." Deep fake videos provide unprecedented customization, targeting, and believability to hostile foreign actors working to spread disinformation.

A video appearing to show House Speaker Nancy Pelosi slurring her words, almost as if she was drunk, spread rapidly on social media in 2019, even being tweeted by former President Trump with the caption, "PELOSI STAMMERS THROUGH NEWS CONFERENCE" (@realDonaldTrump, May 23, 2019). In truth, Pelosi had not slurred her words revise: words, and the video was not a deep fake. Instead, the perpetrators had simply boosted low frequencies in the audio, prompting House Intelligence Chairman Adam Schiff to refer to the effort as a "cheap fake." The episode highlights the threat that deep fake videos pose to our political system and the rapid speed with which they can be seen and spread by millions of people. While news organizations were quick to debunk the video and social media companies worked to stymie its spread, new questions arose about what could be next. What if there were no "real" video to show alongside the fake? What if an accompanying video emerged a few days later showing that Speaker Pelosi had actually been drinking? This is the problem that deep fake videos and audio pose: they create narratives out of whole cloth, with little that can be done to fight them.

This is the problem that deep fake videos and audio pose: they create narratives out of whole cloth, with little that can be done to fight them.

Deep fake videos pose national security risks for the United States when they are spread in a foreign context. A video could be created and spread in another country to show U.S. military personnel engaged in war crimes or the murder of civilians, leading to increased radicalization, violence, and resentment against the U.S. These videos could spread widely before being detected, leaving populations vulnerable to unsuspected threats. Individuals in other countries may also possess lower levels of digital literacy, increasing the likelihood that deep fake videos will be believed. In late 2018 in Gabon, for instance, a video intended to reassure citizens of President Ali Bongo's good health was called a deep fake by his political opponents. They pointed out that his eyes seemed immobile and did not move in sync with his jaw. Outside experts following the controversy said that there was no way to know for sure if the video was a deep fake, but his opponents launched an unsuccessful coup as a result of their belief that it was. Similar tactics could be used around the world to remove U.S.-friendly leaders or cause allies to reconsider their positions toward the U.S.

Deep Fake Detection and Countermeasures

Government agencies, independent researchers, and private companies have created methods and tools able to detect hyper-realistic deep fake photos, video, and audio. The Defense Advanced Research Projects Agency has developed two programs to identify and combat manipulated media: MediFor (media forensics), which assesses the technical integrity of images or videos, and SemaFor (semantic forensics), which assesses semantic issues in manipulated media such as mismatched eye colors and earring placements. Both Google and Facebook have released data sets of deep fake and real videos in hopes of spurring independent innovation of detection methods. Content publishers like Facebook have also imposed greater restrictions in the effort to stop the spread of altered media like deep fakes. However, these restrictions are quite narrow and hard to apply due to their strict requirements about how the manipulated media was created and the intent of the poster, meaning they are likely to be ineffective in fully stopping the threat.

Many deep fakes are low quality and can be easily identified by semantic differences, image quality, and other oddities. In the Katie Jones LinkedIn profile, for example, researchers quickly identified artifacts that made the image look distorted and degraded. Common indicators of deep fake images and videos include skin being overly smooth or lacking details, scaliness or blurriness, flickering, odd head positions, face warping, and unnatural personal patterns of behavior including eye and lip movements.

> ***When these issues are present, it can be easy to debunk fake images and videos; when they are not, more technical solutions are required.***

When these issues are present, it can be easy to debunk fake images and videos; when they are not, more technical solutions are required. Many of these issues have already been solved in the latest GAN frameworks, and CNNs and GANs increasingly make it possible to preserve pose, facial expression and lighting in images and videos, meaning that detection methods will have to be constantly updated to compensate. Independent researchers have created several detection techniques of varying effectiveness and scope; however, more generalized and transferable solutions are still needed.

Deep Fake Photo Detection

One of the most successful methods for detecting deep fake photos relies on artifacts left behind during the creation process. In 2019, Durall et al. used high frequency component analysis to detect artifacts hidden to the human eye indicating that an image may have been manipulated. The team's model achieved 100% accuracy identifying patterns of fakes during supervised learning tests—when a team member offered input and guidance—and 96% accuracy during unsupervised learning tests. Real and fake images have significantly different frequencies that allow them to be classified as either real or fake. While the model struggled to detect lower-resolution deep fakes, this is not a major issue, as these images are less convincing and have less potential to cause harm.

In 2020, Hsu, Zhuang, and Lee used pairwise learning and a common fake feature network to identify deep fake photos. The team's study proposed that by using pairs of images, one real and one fake, they could train their common fake feature network (CFFN) model to identify the most common features of deep fake images. Once the CFFN has been trained to identify the most common features, it can identify whether new images are deep fakes. This method works best on fake face detection, as many of the features across different faces are quite similar, unlike general objects in the world which vary in shape, size, color, and more. Hsu et al. noted that their CFFN may have trouble identifying deep fake images if new generators creating new fakes differ significantly from the generator used to train the CFFN.

Deep Fake Video and Audio Detection

Several video detection methods can be applied to any deep fake video. In 2019, Korshunov and Marcel used two detection techniques to examine the susceptibility of facial recognition software to deep fake face swaps, with varying degrees of success. First, they found that facial recognition software failed up to 95% of the time on deep fake videos, meaning that the software identified the faces in the videos even though they were face-swapped. To combat this issue, they compared an audio-visual approach looking at lip-sync and mouth movements against an image quality technique. They found that the audio-visual approach was highly ineffective, as the deep fake videos accurately matched mouth movements with audio. On the other hand, the image quality technique, which measured signal to noise ratio, blurriness, and other signifiers, was able to identify deep fake videos with more than 90% accuracy.

A similar technique proposed by Güera and Delp uses a recurrent neural network with two components: a

CNN for frame feature extraction and a long short-term memory for temporal sequence analysis. Given an input video, the CNN obtains a set of features for each frame. Next, the features of a consecutive sequence of frames are combined and analyzed by the long short-term memory to produce a likelihood estimate for the probability of a video being a deep fake or not. Their method achieved accuracies greater than 97%, even using less than two seconds of video. This robust and generalized detection method and its ability to achieve high accuracies given low amounts of input will be important to consider in future detection research.

A more specifically tailored deep fake video detection method was proposed by Agarwal et al. to protect world leaders against deep fakes. They extracted data about the facial and head movements from hundreds of hours of footage of U.S. politicians including Barack Obama, Donald Trump, and Bernie Sanders. They found that the specific movements of each individual were quite different, meaning that they could be used to identify that individual. Agarwal et al. then trained a model on both real videos and deep fake videos of each of the leaders and found an average accuracy of 91% across the three main types of deep fake videos. However, their model's accuracy dropped to between 61%–66% for videos where the speaker was not facing the camera. Techniques like this reveal innovative ways that deep fakes can be detected, but it is unlikely that they can be generalized or used to combat deep fakes not targeting famous people.

Implications

On April 26, 2020, the first deep fake targeting the 2020 U.S. election spread widely on Twitter and was retweeted by President Donald Trump. The deep fake, a gif of Vice President Joe Biden raising his eyebrows and rolling his tongue around, originated from a Twitter account called "@SilERabbit" that mainly posted messages in favor of Bernie Sanders, who had dropped out of the Democratic Party's presidential primaries on April 8. While Trump had amplified edited media before, such as the Nancy Pelosi slurring video, this instance was different in that the content was completely fabricated. Journalist David Frum pointed out in *The Atlantic* that Trump's retweet "looks like an experimental test of the rules of social media." It is not clear how the deep fake of Biden appeared in Trump's timeline or if it was sent to him by someone else, but it raises questions if the spread was orchestrated by foreign actors.

This incident may be a sign of a larger shift in how disinformation campaigns since the 2016 election postmortem are being carried out. While the Russian operation to interfere in the 2016 election succeeded in co-opting and influencing news coverage and in engaging many American voters' attention, it failed in that the operation was detected and exposed in great detail. The Russians covered their tracks poorly, leaving behind online transactions, email accounts, correspondence, and other digital identifiers that allowed investigators to paint a detailed picture of the operation and to secure multiple indictments against the perpetrators. In the wake of this exposure, Russia and other foreign actors have sought to increase deniability and believability by outsourcing their operations. In 2019, the *New York Times* reported that "Rather than impersonating Americans as they did in 2016, Russian operatives are working to get Americans to repeat disinformation."

While the goals and content have remained constant, the tactics have changed, making it harder to track the origin of disinformation and the perpetrators behind it.

Evidence of the shift in Russian tactics has emerged in Africa over the past year. In late October 2019, Facebook removed three networks of accounts that had been spreading disinformation in Mozambique, Cameroon, Sudan, and Libya. These accounts were linked to Yevgeny Prigozhin, who the U.S. indicted for meddling in the 2016 election. A 2020 CNN investigation found that Russian operatives linked to the Internet Research Agency have outsourced the actual running of accounts and posting to workers in nations like Ghana and Nigeria. These Russian-backed trolls have posted content targeted towards Americans to incite racial tensions and social unrest. While the goals and content have remained constant, the tactics have changed, making it harder to track the origin of disinformation and the perpetrators behind it.

While deep fakes today are usually easy to spot, they may not be in the future. Research into detection algorithms must at least match the development of creation algorithms. In turn, social media companies like Twitter and Facebook will need to employ these techniques at scale on their platforms. Major news organizations and other groups focused on fighting disinformation and providing transparency in technology will also need to adopt them.

The emergence of deep fakes presents many immediate challenges, but the broader issue is the continuing and deepening challenge to truth in our discourse. The U.S. is already incredibly divided by partisan rhetoric and media organizations that spread tensions across the political spectrum. Foreign actors further inflame these tensions, leading to greater distrust in institutions and a lack of regard for the truth. Citron and Chesney refer to this deepening spiral as the "liar's dividend," in which citizens' growing awareness of deep fakes makes it increasingly easy to question the truth in any situation. As the public becomes more aware that deep fakes could be anywhere, they "may have difficulty believing what their eyes or ears are telling them—even when the information is real."

It is unclear whether this new reality of disinformation, charged rhetoric, and increasing skepticism is a fleeting element of the moment or if it is here to stay.

Americans have rarely vested full faith in their government, its institutions, and the media who report on both, but recent shifts in information and discourse have been rapid and startling. Accusations of "fake news," a term all but unheard of before 2016, are levied against all critical reporting by those who hold positions of power, no matter how valid. It is unclear whether this new reality of disinformation, charged rhetoric, and increased skepticism is a fleeting element of the moment or if it is here to stay, but deep fakes are certainly accelerating their influence on American discourse.

The fight for truth in American discourse faces a grim future. At the same time that foreign influence campaigns are becoming less expensive to operate and more successful in their reach and effect, the algorithms and programs used to create deep fakes are advancing much more rapidly than detection algorithms, regulations, laws, and societal demand for change. Russia may have invented the playbook in 2016 for successful online disinformation campaigns, but other nations and groups have been quick to adopt Russia's strategies. The Intelligence Community has already assessed that Iran and China have ramped up their election interference schemes, but disinformation does not stop at our elections.[1] Socio-political developments like the COVID-19 pandemic provide opportunities for malicious actors to spread disinformation and increase political tensions and polarization in the U.S. Deep fakes increase the potential damage of disinformation campaigns in too many imaginable ways, providing unprecedented believability to complete fabrications.

Author's Note

Nick Dunard

Nick Dunard ('21) graduated with a bachelor's degree in Intelligence Analysis with a minor in Political Science. As an undergraduate, Nick studied the intersections between national security, emerging technologies, and politics. He will continue his studies at the Catholic University of America in the Columbus School of Law and hopes to work as a lawyer in national security ethics and oversight.

Bibliography

Agarwal, Shruti, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. "Protecting World Leaders Against Deep Fakes." Paper presented at IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, California, June 16-19, 2019. https://farid.berkeley.edu/downloads/publications/cvpr19/cvpr19a.pdf.

Alba, Davey, and Sheera Frenkel. "Russia Tests New Disinformation Tactics in Africa to Expand Influence." *New York Times*, October 30, 2019. https://www.nytimes.com/2019/10/30/technology/russia-facebook-disinformation-africa.html.

BBC News. "Mother 'Used Deepfake to Frame Cheerleading Rivals.'" March 15, 2021. https://www.bbc.com/news/technology-56404038.

Bickert, Monika. "Enforcing Against Manipulated Media." Facebook, January 6, 2020. https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/.

Breland, Ali. "The Bizarre and Terrifying Case of the "Deepfake" Video That Helped Bring an African Nation to the Brink." *Mother Jones*, March 15, 2019. https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/.

Breuninger, Kevin, and Amanda Macias. "Russia and Iran Tried to Interfere with 2020 Election, U.S. Intelligence Agencies Say." *CNBC*. March 16, 2021. https://www.cnbc.com/2021/03/16/russia-and-iran-tried-to-interfere-with-2020-election-us-intelligence-agencies-say.html.

Chesney, Bobby, and Danielle Citron. "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security." *California Law Review* 107, no. 6 (2019): 1753-1820. https://heinonline.org/HOL/Page?handle=hein.

journals/calr107&div=51&g_sent=1&casa_token=&collection=journals.

Coats, Daniel R. *Worldwide Threat Assessment of the U.S. Intelligence Community.* Washington, D.C., 2019. https://www.dni.gov/files/ODNI/documents/2019-ATA-SFR---SSCI.pdf.

Dufour, Nick, and Andrew Gully. "Contributing Data to Deepfake Detection Research." *Google AI Blog*, September 24, 2019. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html

Durall, Ricard, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. "Unmasking DeepFakes with Simple Features." *arXiv* (2019). https://arxiv.org/pdf/1911.00686.pdf.

End Revenge Porn. "Revenge Porn Statistics." *Cyber Civil Rights Initiative*, n.d. https://www.cybercivilrights.org/wp-content/uploads/2014/12/RPStatistics.pdf.

Engler, Alex. "Fighting Deepfakes When Detection Fails." *Brookings Institution,* November 14, 2019. https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/.

Fabian, Jordan. "US Warns of 'Ongoing' Election Interference by Russia, China, Iran." *The Hill*, October 19, 2018. https://thehill.com/policy/national-security/412292 -us-warns-of-ongoing-election-interference-by-russia-china-iran.

Frum, David. "The Very Real Threat of Trump's Deep-Fake." *The Atlantic,* April 27, 2020. https://www.theatlantic.com/ideas/archive/2020/04/trumps-first-deepfake/610750/.

Geron, Aurélien. *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow.* Sebastopol: O'Reilly Media, Inc., 2019.

Giles, Martin. "The GANfather: The Man Who's Given Machines the Gift of Imagination." *MIT Technology Review,* February 21, 2018. https://www.technologyreview.com/2018/02/21/145289/the-ganfather-the-man-whosgiven-machines-the-gift-of-imagination/.

Goldman, Adam, Barnes, Julian, Haberman, Maggie, and Fandos Nicholas. "Lawmakers Are Warned That Russia Is Meddling to Re-Elect Trump." *New York Times*, Febuary 20, 2020. https://www.nytimes.com/2020/02/20/us/politics/russian-interference-trump-democrats.html.

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Nets." *arXiv* (2014): 1-9. https://arxiv.org/pdf/1406.2661.pdf.

Güera, David, and Edward J. Delp. "Deepfake Video Detection Using Recurrent Neural Networks." Paper presented at IEEE International Conference on Advanced Video and Signal Based Surveillance. Auckland, New Zealand, 2018. https://doi.org/10.1109/AVSS.2018.8639163.

Hao, Karen. "Deepfake Porn Is Ruining Women's Lives. Now the Law May Finally Ban It." *MIT Technology Review,* February 12, 2021. https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/.

Hsu, Chih-Chung, Yi-Xiu Zhuang, and Chia-Yen Lee. "Deep Fake Image Detection Based on Pairwise Learning." *Applied Sciences* 10, no. 1 (2020): 1-14. https://doi.org/10.3390/app10010370.

Hubel, David H. "Single Unit Activity in Striate Cortex of Unrestrained Cats." *The Journal of Physiology* 147, no. 2 (1959): 226-238. https://doi.org/10.1113/jphysiol.1959.sp006238.

Hubel, David H., and Torsten N. Wiesel. "Receptive Fields of Single Neurons in Cat's Striate Cortex." *The Journal of Physiology* 148, no. 3 (1959): 574-591. https://doi.org/10.1113/jphysiol.1959.sp006308.

Karras, Tero, Samuli Laine, and Timo Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks." *arXiv* (2019): 1-12. https://arxiv.org/abs/1812.04948?amp=1.

Korshunov, Pavel, and Sebastien Marcel. "Vulnerability Assessment and Detection of Deepfake Videos." Paper presented at IAPR International Conference on Biometrics, Crete, Greece, 2019. https://doi.org/10.1109/ICB45273.2019.8987375.

Mao, Xudong, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. "Least Squares Generative Adversarial Networks." Paper presented at IEEE International Conference on Computer Vision, Venice, Italy, 2017. https://doi.org/10.1109/ICCV.2017.304.

Mack, David. "This PSA about Fake News from Barack Obama Is Not What It Appears." *BuzzFeed News,* April 17, 2018. https://www.buzzfeednews.com/article/davidmack/obama-fake-news-jordan-peele-psavideo-buzzfeed.

Maksutov, Artem A., Viacheslav O. Morozov, Aleksander A. Lavrenov, and Alexander S. Smirnov. “Methods of Deepfake Detection Based on Machine Learning.” Paper presented at IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, St. Petersburg and Moscow, Russia, 2020, 408-411. https://doi.org/10.1109/EIConRus49466.2020.9039057.

Mervosh, Sarah. “Distorted Videos of Nancy Pelosi Spread on Facebook and Twitter, Helped by Trump.” *New York Times,* May 24, 2019. https://www.nytimes.com/2019/05/24/us/politics/pelosi-doctored-video.html.

Mueller, Robert S., III. *Report on the Investigation into Russian Interference in the 2016 Presidential Election.* Washington, D.C., 2019. https://www.justice.gov/storage/report.pdf.

National Institute of Neurological Disorders and Stroke. “Brain Basics: The Life and Death of a Neuron,” *National Institutes of Health*, last modified December 16, 2019, https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/life-and-death-neuron.

Nguyen, Thanh Thi, Coung M. Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, and Saeid Nahavandi. “Deep Learning for Deepfakes Creation and Detection.” *arXiv* (2019): 1-16. https://arxiv.org/pdf/1909.11573.pdf.

O’Sullivan, Donie. “Congress to Investigate Deepfakes as Doctored Pelosi Video Causes Stir.” *CNN*, June 4, 2019, https://www.cnn.com/2019/06/04/politics/house-intelligence-committee-deepfakes-threats-hearing.

Radford, Alec, Luke Metz, and Soumith Chintala. “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks.” *arXiv* (2016): 1-16. https://arxiv.org/abs/1511.06434.

Satter, Raphael. “Experts: Spy Used AI-Generated Face to Connect with Targets.” *Associated Press,* June 13, 2019. https://apnews.com/bc2f19097a4c4fffaa00de6770b8a60d.

Sayler, Kelley M., and Laurie A. Harris. “Deep Fakes and National Security.” *Congressional Research Service*, October 14, 2019. https://fas.org/sgp/crs/natsec/IF11333.pdf.

Turek, Matt. “Media Forensics (MediFor).” *Defense Advanced Research Projects Agency*, n.d. https://www.darpa.mil/program/media-forensics.

———. “Semantic Forensics (SemaFor).” *Defense Advanced Research Projects Agency,* n.d. https://www.darpa.mil/program/semantic-forensics.

Ward, Clarissa, Katie Polglase, Sebastian Shukla, Gianluca Mezzofiore, and Tim Lister. “Russian Election Meddling Is Back—Via Ghana and Nigeria—and in Your Feeds.” *CNN*, April 11, 2020. https://www.cnn.com/2020/03/12/world/russia-ghana-troll-farms-2020-ward/index.html.

Wong, Edward, Matthew Rosenberg, and Julian E. Barnes. “Chinese Agents Helped Spread Messages That Sowed Virus Panic in U.S., Officials Say.” *New York Times,* April 22, 2020. https://www.nytimes.com/2020/04/22/us/politics/coronavirus-china disinformation.html.

Endnotes

[1] Aurélien Geron, *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow* (Sebastopol: O’Reilly Media, Inc., 2019), 279.

[2] National Institute of Neurological Disorders and Stroke, “Brain Basics: The Life and Death of a Neuron,” *National Institute of Neurological Disorders and Stroke,* last modified December 16, 2019, https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Life-and-Death-Neuron.

[3] Geron, *Hands-On Machine Learning,* 281.

[4] David H. Hubel and Torsten N. Wiesel, “Receptive Fields of Single Neurons in Cat’s Striate Cortex,” *The Journal of Physiology* 148, no. 3 (1959): 574-591, https://doi.org/10.1113/jphysiol.1959.sp006308; David H. Hubel, “Single Unit Activity in Striate Cortex of Unrestrained Cats,” *The Journal of Physiology* 147, no. 2 (1959): 226-238, https://doi.org/10.1113/jphysiol.1959.sp006238.

[5] Geron, *Hands-On Machine Learning*, 448.

[6] Ian J. Goodfellow et al., “Generative Adversarial Nets,” *arXiv* (2014): 1, https://arxiv.org/pdf/1406.2661.pdf.

[7] Martin Giles, “The GANfather: The Man Who's Given Machines the Gift of Imagination,” *MIT Technology Review*, February 21, 2018, https://www.technologyreview.com/2018/02/21/145289/the-ganfather-the-man-whos-given-machines-the-gift-of-imagination/.

[8] Alec Radford, Luke Metz, and Soumith Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,” *arXiv* (2016): 1, https://arxiv.org/abs/1511.06434.

[9] Xudong Mao et. al., “Least Squares Generative Adversarial Networks" (paper, International Conference on Computer Vision, Venice, Italy, 2017), http://openaccess.thecvf.com/content_ICCV_2017/papers/Mao_Least_Squares_Generative_ICCV_2017_paper.pdf.

[10] Tero Karras, Samuli Laine, and Timo Aila, “A Style-Based Generator Architecture for Generative Adversarial Networks,” *arXiv* (2019): 1, https://arxiv.org/abs/1812.04948?amp=1.

[11] Shruti Agarwal et al., “Protecting World Leaders Against Deep Fakes" (paper, at IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, California, June 16-19, 2019), 1, https://farid.berkeley.edu/downloads/publications/cvpr19/cvpr19a.pdf.

[12] David Mack, “This PSA about Fake News from Barack Obama Is Not What It Appears,” *Buzzfeed News*, April 17, 2018, https://www.buzzfeednews.com/article/davidmack/obama-fake-news-jordan-peele-psa-video-buzzfeed.

[13] Robert S. Mueller III. *Report on the Investigation into Russian Interference in the 2016 Presidential Election,* (Washington, D.C., 2019), 1, https://www.justice.gov/storage/report.pdf.

[14] Daniel R Coats. *Worldwide Threat Assessment*, (Washington, D.C., 2019), 7, https://www.dni.gov/files/ODNI/documents/2019-ATA-SFR---SSCI.pdf 7; Kevin Breuninger and Amanda Macias, “Russia and Iran Tried to Interfere with 2020 Election, U.S. Intelligence Agencies Say,” *CNBC*, March 16, 2021, https://www.cnbc.com/2021/03/16/russia-and-iran-tried-to-interfere-with-2020-election-us-intelligence-agencies-say.html.

[15] Edward Wong, Matthew Rosenberg, and Julian E. Barnes, “Chinese Agents Helped Spread Messages That Sowed Virus Panic in U.S., Officials Say,” *New York Times*, April 22, 2020, https://www.nytimes.com/2020/04/22/us/politics/coronavirus-china-disinformation.html.

[16] Coats. *Worldwide Threat Assessment of the U.S. Intelligence Community*, 7.

[17] Kelley M. Sayler and Laurie A. Harris, “Deep Fakes and National Security,” *Congressional Research Service*, October 14, 2019, https://fas.org/sgp/crs/natsec/IF11333.pdf.

[18] Raphael Satter, “Experts: Spy Used AI-Generated Face to Connect with Targets,” Associated Press, June 13, 2019, https://apnews.com/bc2f19097a4c4fffaa00de6770b8a60d.

[19] Satter, “Experts.”

[20] Bobby Chesney and Danielle Keats Citron, “Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security,” *California Law Review* 107, no. 6 (December 2019): 1757, https://heinonline.org/HOLPage?handle=hein.journals/calr107&div=51&g_sent=1&casa_token=&collection=-journals.

[21] Karen Hao, “Deepfake Porn Is Ruining Women’s Lives. Now the Law May Finally Ban It,” *MIT Technology Review*, February 12, 2021, https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/.

[22] End Revenge Porn, “Revenge Porn Statistics,” *Cyber Civil Rights Initiative*, n.d., https://www.cybercivilrights.org/wp-content/uploads/2014/12/RPStatistics.pdf.

[23] Hao, "Deepfake Porn Is Ruining Women's Lives."

[24] BBC News, “Mother ‘Used Deepfake to Frame Cheerleading Rivals,’” March 15, 2021, https://www.bbc.com/news/technology-56404038.

[25] Chesney and Citron, “Deep Fakes,” 1776.

[26] Mueller III, *Report on the Investigation into Russian Interference*, 4.

[27] Sarah Mervosh, “Distorted Videos of Nancy Pelosi Spread on Facebook and Twitter, Helped by Trump,” *New York Times*, May 24, 2019, https://www.nytimes.com/2019/05/24/us/politics/pelosi-doctored-video.html.

[28] Mervosh, “Distorted Videos.; Donie O’Sullivan, “Congress to Investigate Deepfakes as Doctored Pelosi Video Causes Stir,” *CNN*, June 4, 2019, https://www.cnn.com/2019/06/04/politics/house-intelligence-committee-deepfakes-threats-hearing/index.html.

[29] Ali Breland, “The Bizarre and Terrifying Case of the “Deepfake” Video That Helped Bring an African Nation to the Brink,” *Mother Jones*, March 15, 2019, https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/.

[30] Breland, “The Bizarre and Terrifying Case.”

[31] Matt Turek, “Media Forensics (MediFor),” *Defense Advanced Research Projects Agency,* n.d., https://www.darpa.mil/program/media-forensics.; Matt Turek, “Semantic Forensics (SemaFor),” *Defense Advanced Research Projects Agency,* n.d., https://www.darpa.mil/program/semantic-forensics.

[32] Monika Bickert, “Enforcing Against Manipulated Media,” Facebook, January 6, 2020, https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/.

[33] Artem A. Maksutov, Viacheslav O. Morozov, Aleksander A. Lavrenov, and Alexander S. Smirnov, “Methods of Deepfake Detection Based on Machine Learning" (paper, IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, St. Petersburg and Moscow, Russia, 2020), 410, https://doi.org/10.1109/EIConRus49466.2020.9039057.

[34] Thanh Thi Nguyen et al., “Deep Learning for Deepfakes Creation and Detection,” *arXiv* (2019): 4, https://arxiv.org/pdf/1909.11573.pdf.

[35] Ricard Durall et al., “Unmasking DeepFakes with Simple Features,” *arXiv* (2020), https://arxiv.orgpdf/1911.00686.pdf.

[36] Chih-Chung Hsu, Yi-Xiu Zhuang, and Chia-Yen Lee, “Deep Fake Image Detection Based on Pairwise Learning,” *Applied Sciences* 10, no. 1 (2020): 3-4.

[37] Hsu, Zhuang, and Lee, “Deep Fake Image Detection,” 6.

[38] Pavel Korshunov and Sebastien Marcel, “Vulnerability Assessment and Detection of Deepfake Videos" (paper, IAPR International Conference on Biometrics, Crete, Greece, 2019): 5, https://doi.org/10.1109/ICB45273.2019.8987375.

[39] Korshunov and Marcel, “Vulnerability Assessment and Detection,” 5.

[40] David Güera and Edward J. Delp, “Deepfake Video Detection Using Recurrent Neural Networks" (paper, IEEE International Conference on AdvancedVideo and Signal Based Surveillance, Auckland, New Zealand, 2018): 3, https://doi.org/10.1109/AVSS.2018.8639163.

[41] Güera and Delp, “Deepfake Video Detection,” 5.

[42] Agarwal et al., “Protecting World Leaders Against Deep Fakes,” 2.

[43] David Frum, “The Very Real Threat of Trump’s Deepfake,” *The Atlantic*, April 27, 2020, https://www.theatlantic.com/ideas/archive/2020/04/trumps-first-deepfake/610750/.

[44] Mueller III, *Report on the Investigation into Russian Interference,* 4-7.

[45] Goldman et al., “Lawmakers Are Warned That Russia Is Meddling to Re-Elect Trump.” *New York Times*, Febuary 20, 2020. https://www.nytimes.com/2020/02/20/us/politics/russian-interference-trump-democrats.html.

[46] Davey Alba and Sheera Frenkel, “Russia Tests New Disinformation Tactics in Africa to Expand Influence,” *New York Times*, October 30, 2019, https://www.nytimes.com/2019/10/30/technology/russia-facebook-disinfomation-africa.html.

[47] Alba and Frenkel, “Russia Tests New Disinformation Tactics."

[48] Clarissa Ward et al., “Russian Election Meddling Is Back – Via Ghana and Nigeria – and in Your Feeds,” *CNN*, April 11, 2020.

[49] Ward et al., “Russian Election Meddling Is Back.”

[50] Chesney and Citron, “Deep Fakes: A Looming Challenge,” 1785.

[51] Chesney and Citron, “Deep Fakes: A Looming Challenge,” 1786.

[52] Jordan Fabian, “US Warns of ‘Ongoing’ Election Interference by Russia, China, Iran,” *The Hill*, October 19, 2018, https://thehill.com/policy/national-security/412292-us-warns-of-ongoing-election-interference-by-russia china-iran.