CSET CENTER for SECURITY and EMERGING TECHNOLOGY

JULY 2020

Deepfakes

A Grounded Threat Assessment

AUTHOR
Tim Hwang

CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

Established in January 2019, the Center for Security and Emerging Technology (CSET) at Georgetown's Walsh School of Foreign Service is a research organization focused on studying the security impacts of emerging technologies, supporting academic work in security and technology studies, and delivering nonpartisan analysis to the policy community. CSET aims to prepare a generation of policymakers, analysts, and diplomats to address the challenges and opportunities of emerging technologies. During its first two years, CSET will focus on the effects of progress in artificial intelligence and advanced computing.

CSET.GEORGETOWN.EDU | CSET@GEORGETOWN.EDU

JULY 2020

Deepfakes

A Grounded Threat Assessment

AUTHOR
Tim Hwang

ACKNOWLEDGMENTS

The author would like to thank James Dunham, Melissa Flagg, Tina Huang, Matt Mahoney, Maura McCarthy, Igor Mikolic-Torreira, Dewey Murdick, Osonde Osoba, Tim Rudner, Rex Sorgatz, Helen Toner, Alexandra Vreeman, Clint Watts, and Lynne Weil for their invaluable feedback on earlier drafts of this report.

PRINT AND ELECTRONIC DISTRIBUTION RIGHTS

Cover photo: pickup/AdobeStock

Contents

Executive Summary

Researchers have used machine learning (ML) in recent years to generate highly realistic fake images and videos known as "deepfakes." Artists, pranksters, and many others have subsequently used these techniques to create a growing collection of audio and video depicting high-profile leaders, such as Donald Trump, Barack Obama, and Vladimir Putin, saying things they never did. This trend has driven fears within the national security community that recent advances in ML will enhance the effectiveness of malicious media manipulation efforts like those Russia launched during the 2016 U.S. presidential election.

These concerns have drawn attention to the disinformation risks ML poses, but key questions remain unanswered. How rapidly is the technology for synthetic media advancing, and what are reasonable expectations around the commoditization of these tools? Why would a disinformation campaign choose deepfakes over more crudely made fake content that is sometimes equally as effective? What kinds of actors are likely to adopt these advances for malicious ends? How will they use them? Policymakers and analysts often lack concrete guidance in developing policies to address these risks.

This paper examines the technical literature on deepfakes to assess the threat they pose. It draws two conclusions. First, the malicious use of crudely generated deepfakes will become easier with time as the technology commodifies. Yet the current state of deepfake detection suggests that these fakes can be kept largely at bay.

Second, tailored deepfakes produced by technically sophisticated actors will represent the greater threat over time. Even moderately resourced

campaigns can access the requisite ingredients for generating a custom deepfake. However, factors such as the need to avoid attribution, the time needed to train an ML model, and the availability of data will constrain how sophisticated actors use tailored deepfakes in practice.

Based on this assessment, the paper makes four recommendations:

- **Build a Deepfake "Zoo":** Identifying deepfakes relies on rapid access to examples of synthetic media that can be used to improve detection algorithms. Platforms, researchers, and companies should invest in the creation of a deepfake "zoo" that aggregates and makes freely available datasets of synthetic media as they appear online.
- **Encourage Better Capabilities Tracking:** The technical literature around ML provides critical insight into how disinformation actors will likely use deepfakes in their operations, and the limitations they might face in doing so. However, inconsistent documentation practices among researchers hinders this analysis. Research communities, funding organizations, and academic publishers should work toward developing common standards for reporting progress in generative models.
- **Commodify Detection:** Broadly distributing detection technology can inhibit the effectiveness of deepfakes. Government agencies and philanthropic organizations should distribute grants to help translate research findings in deepfake detection into user-friendly apps for analyzing media. Regular training sessions for journalists and professions likely to be targeted by these types of techniques may also limit the extent to which members of the public are duped.
- **Proliferate Radioactive Data:** Recent research has shown that datasets can be made "radioactive." ML systems trained on this kind of data generate synthetic media that can be easily identified. Stakeholders should actively encourage the "radioactive" marking of public datasets likely to train deep generative models. This would significantly lower the costs of detection for deepfakes generated by commodified tools. It would also force more sophisticated disinformation actors to source their own datasets to avoid detection.

Introduction

Researchers have used machine learning (ML) in recent years to generate highly realistic fake images and videos known as "deepfakes." Artists, pranksters, and many others have subsequently used these techniques to create a growing collection of audio and video depicting high-profile leaders, such as Donald Trump, Barack Obama, and Vladimir Putin, saying things that they never did.[1]

This trend has driven fear within the national security community that states, political parties, and malicious actors will leverage recent advances in ML to enhance the effectiveness of their media manipulation efforts. A number of cases in which these technologies have been weaponized to generate synthetic non-consensual pornography bolster these concerns.[2] Policymakers and analysts in the national security community foresee the integration of cutting-edge ML technologies into large-scale information warfare efforts like the one Russia launched during the 2016 U.S. presidential election.[3]

Public discourse around this risk has centered on a broad, amorphous fear that synthetic media will erode the ability to discern the real from the fake. One New York Times op-ed writer encapsulated the framing in 2019: "Deepfakes Are Coming. We Can No Longer Believe What We See."[4] Similarly broad concerns have shaped government discourse on these issues. After listing some possible malicious uses at a hearing on the matter, the chairman of a congressional committee acknowledged, "one does not need any great imagination to envision even more nightmarish scenarios that would leave the government, the media, and the public struggling to discern what is real and what is fake."[5]

These generalized fears have galvanized attention around the technology's disinformation risks. However, a lack of answers to key questions has left policymakers and analysts without clear guidance in developing policies to address these risks. How rapidly is the technology for synthetic media advancing, and what are reasonable expectations around the commoditization of these tools? Why would a disinformation campaign choose to distribute deepfakes instead of more crudely made fake content sometimes equally as effective? What kinds of actors are likely to adopt these advances for malicious ends? How will they use them?

Through a review of technical literature in the field of ML, this paper answers these questions to offer an assessment of the threat deepfakes pose.

In this paper, the term "deepfakes" refers to the broad scope of synthetic images, video, and audio generated through recent breakthroughs in the field of ML, specifically in deep learning. This term is inclusive of ML techniques that seek to modify some aspect of an existing piece of media, or to generate entirely new content. While this paper emphasizes advances in neural networks, its analysis is relevant for other methods in the broader field of ML. The term "deepfakes" excludes the wide range of techniques for manipulating media without the use of ML, including many existing tools for "cutting and pasting" objects from one image to another.

This paper is organized in three parts. The first one lays out a framework for assessing the potential impact of deepfakes in the media manipulation domain. The second part reviews the current state of the art, both in the creation of synthetic media and in the detection of media generated by these techniques. The third part brings together various trends in the research field to predict how disinformation campaigns might use these synthetic media techniques going forward. This paper concludes by offering policy recommendations based on this analysis.

Malicious actors might enhance their influence operations with ML in many ways. The focus of this paper is squarely on synthetic images, audio, and video, currently most prominent in the public eye and of concern to national security scholars and policymakers. This paper excludes a review of technologies like conversational systems, which might be used to drive realistic bot identities on social media. It also eschews a close examination of ML-driven text generation, which could quickly populate "fake news" websites and forge written documents. However, the general framework offered in this paper could also be applied to understand the impact of these areas of research on disinformation threats; it would be valuable to do so going forward.

1 A Framework for Threat Assessment

ML has a wide range of possible applications. Computer vision—the subfield of ML that gives computers the ability to process and comprehend imagery—can be used to take on tasks as disparate as drone piloting and understanding gender bias in cinema.[6] Within a single application, ML can be configured to handle a range of different subtasks. Consider the application of ML in the context of an automobile. ML might be applied to assist in navigating the vehicle to its destination, avoiding unexpected obstacles on the road, modeling driver behavior for marketing purposes, some subset of these tasks, or none.

Given the wide range of potential ML applications, asking whether ML *could* be applied in a specific domain or to a particular problem often proves unhelpful. It is always possible to construct a scenario in which the technology might be used, for better or for worse. But this abstract speculation does little to reveal how ML will be used in practice: the fact a technology can do something does not mean it will be put to that purpose.

A deeper understanding of individual and organizational incentives is needed to understand how ML will actually be used. These incentives will determine if ML is an attractive technology to adopt and how it will be implemented.

This section of the paper lays out the incentives shaping whether and how disinformation campaigns will use deepfakes in their operations. First, it summarizes what is currently known about how disinformation campaigns spread false information, arguing that influence operations weigh deciding factors like cost and impact when choosing what kinds of content to distribute. It then examines these deciding factors in the context of deepfakes, laying out three key drivers influencing how disinformation campaigns use the technology in practice.

PROPAGANDISTS ARE PRAGMATISTS

Online propagandists are pragmatists. They seek to wield the greatest degree of social and political influence at the lowest possible cost. Currently, these incentives lead disinformation campaigns to distribute crude photo edits and copied images, rather than high-production hoaxes.

The effort by the Internet Research Agency—an organization backed by the Russian government—to interfere in the 2016 U.S. presidential election is a prototypical case.[7] The IRA's content emphasized scale over quality. It posted large numbers of images and videos that were cheaply crafted or taken from elsewhere on the web.[8] Numerous other online influence campaigns have exhibited a similar operational posture.[9]

Perpetrators of disinformation campaigns have not made a significant investment in crafting high-quality hoaxes, with or without ML. One explanation is that cheaper techniques of content production have already proven successful; crudely editing image and video, as well as appropriating content from elsewhere and attaching a false description, are effective tactics for spreading false narratives throughout the web.[10] Online influence campaigns may not even need a faked image or video to push a message. Images featuring slogans, informational graphics, or purely symbolic imagery were some of the most widely shared content of the 2016 election interference campaign.[11]

The fact that disinformation campaigns rely on cheap, rough-and-ready ways of producing content suggests that practical considerations figure into the types of content they spread. There is no need to spend additional resources creating an elaborate fake video when simply copying an image from elsewhere and misleadingly captioning it will achieve the same impact.

But the past may not be a good guide to the future. ML technologies are rapidly evolving in ways that will change whether or not disinformation campaigns invest in creating high-quality fakes.

KEY DRIVERS SHAPING DEEPFAKE USE

What practical considerations must disinformation campaigns weigh in deciding whether to use deepfakes in their operations? Three significant factors will shape their use of the technology: the persuasive capacity of ML-driven synthetic media, the operational requirements of implementing the technology, and the novel risks of detection raised by using deepfakes.

Benefit: Persuasive Capacity

Deepfakes offer the online propagandist a unique opportunity to create hoaxed content. ML-driven fakery can produce strikingly realistic depictions of individuals

and situations. Importantly, deepfakes can reproduce various subtle details—such as convincing facial tics or realistic shadows for a fake object pasted into an image—that make it challenging to identify an image or video as a hoax. At the very least, such fakes may sow sufficient doubt about a target individual or situation to create confusion and suspicion.[12]

But this increased realism does not necessarily make deepfakes a clear-cut choice for disinformation campaigns. A perfect simulation of a voice or facial movement does not inherently mean a hoax will be believed and shared.

The internet is rife with examples of crudely produced fakes that are widely circulated and taken as true. Consider the 2019 video of Speaker of the House Nancy Pelosi that spread extensively through social media, purporting to show Pelosi either drunk or suffering from some kind of mental deterioration. No ML was used in this case. The video was produced simply by slowing down a real video of Pelosi speaking at an event.[13]

Malicious actors clearly do not need to achieve visual realism for their hoaxes to succeed. Instead, playing on "motivated reasoning"—the tendency to accept information confirming pre-existing prejudices—may be a more important factor in the success of a hoax image or video, rendering the enhanced visuals possible through ML irrelevant for the propagandist. This makes deepfakes a less attractive method for spreading false narratives, particularly when weighing the costs and risks of using the technology.

Cost: Operational Requirements

Disinformation campaigns cannot adopt ML without incurring certain operational costs. As a general matter, the creation of high-performance AI systems requires access to a sufficient training data (enabling a machine to learn how to accomplish a given task) and computational power (the hardware needed to execute the training process). Depending on what the deepfake depicts, there may be significant expense in acquiring the training data, structuring it properly, and running the training process. A perpetrator may also need specialized expertise, making deepfakes a more expensive option than the crude media editing and copying currently characterizing online disinformation operations.

On the other hand, deepfake technologies are increasingly integrated into software platforms that do not require special technical expertise. Easy-to-use, ML-driven software that facilitates a "face swap"—removing one face from an image or video and inserting another—is increasingly available for users with no technical expertise.[14] Other routine transformations of images and video powered by ML are likely to follow. This trend toward democratization may reduce or effectively eliminate many of the operational costs otherwise making deepfakes an unattractive option for disinformation perpetrators.

Risks: Algorithmic Detection

Influence operations prefer to avoid public exposure. A political campaign caught using underhanded methods to manipulate online discourse may face public censure and legal action. A country found to be doing the same may face retaliation and sanction. Social media companies can "deplatform" an influence campaign when discovered, deleting accounts and otherwise hindering malicious actors' access to users.[15] At the very least, widespread knowledge about an ongoing influence effort may put an otherwise unsuspecting population on guard.

Online influence operations may increase their risk of exposure by using deepfakes. The ML models for producing deepfakes can leave suspicious distortions in images, audio, and video that are often consistent across content distributed by an influence campaign. Deepfakes may therefore contain a kind of "fingerprint," allowing investigators to link together all media originating from a given disinformation campaign. Investigators, in turn, can trace the campaign to a specific source and alert the public.

This poses a problem for influence campaigns because their success depends on wide distribution of their content through intermediaries, such as Facebook, Twitter, and YouTube. As fears over deepfakes have escalated, these platforms have created new policies prohibiting the use of certain kinds of synthetic media.[16] These policies will use detection algorithms for enforcement, given the massive scale of content uploaded and shared on social media. By choosing to distribute deepfakes, influence operations run the risk of their messaging being quickly taken down or flagged as suspicious on these platforms.

These increased risks of exposure and detection may make deepfakes a less attractive means of spreading false narratives than existing methods. Manually copying content from many sources and editing media as needed may avoid the consistent "fingerprints" left by ML models. Diversity of content makes it more challenging to build algorithms that identify disinformation in a vast stream of social media activity. Similarly, appropriating content and re-contextualizing it in a false or misleading way—a tactic seen in many disinformation efforts—may present no consistent patterns that allow for effective, automated filtration.

The adoption of deepfakes for disinformation purposes will therefore depend on more than the costs of producing this content and its likely impact on the target audience. It will also depend on the speed of improvement in deepfake detection and the adoption of detection technologies by online platforms, governments, and everyday users.

APPLYING THE ANALYSIS

The decision of malicious actors to adopt deepfakes for disinformation efforts will be influenced by the nature of ML technology. Three key factors determine whether and how online influence operations will use deepfakes:

- **What can be depicted in a deepfake.** Disinformation actors will adopt ML only if it creates synthetic media likely to shape public perceptions or cast doubt in the minds of a target audience.
- **The computational, human, and data requirements of generating deepfakes.** High costs of production will make deepfakes less attractive relative to manual methods, while low production costs will make them more attractive.
- **The effectiveness of detection systems.** The ability to detect deepfakes at low cost makes ML less attractive to disinformation actors, while ineffectual or high costs to detection make it more attractive.

The next section of this paper considers these variables, examining the costs of deepfake generation, the scope of what media can be generated through ML methods, and the state of play in deepfake detection. Based on this analysis, the final section offers predictions about the evolution of the ML threat in media manipulation.

2 The State of Play

The state of the ML field will define the persuasive capacity, operational requirements, and detection risks of deepfakes. This section examines the current state of play in deepfake creation and deepfake detection—the most cutting-edge results seen in the lab, as well as trends in the adaptation of these research findings into practical tools for use in the field. Reviewing the research literature provides vital clues as to the actual threat from deepfakes and how malicious actors will use this technology to spread disinformation.

DEEPFAKE CREATION

Deepfakes must meet two criteria in order for online influence campaigns to use them. First, the operational costs of producing a deepfake—buying hardware, acquiring data, and hiring expert engineers—must not be overly onerous. Second, deep generative models must be able to successfully produce the faked media an influence campaign seeks to distribute.

This section explores the recent technical literature in ML in order to assess deepfakes against these two criteria. To provide context for this review, this section first explains how deepfakes are generated. It then looks at the current state of the art in using ML to produce synthetic media.

This is by necessity a limited review. There are no standardized practices for documenting the training data, models, and hardware used in an experiment. This makes it difficult to extract a comprehensive picture of the resource requirements for generating a state-of-the-art deepfake from the research literature. It is similarly challenging to determine the evolving trade-off between cost and quality in this domain: comparing images

different deep generative networks produce is subjective, and researchers have criticized the standard performance measurements in recent years.[17] Moreover, differences in training datasets, hardware, and learning architectures across research publications make rigorous comparisons challenging.[18] In spite of these issues, recent research papers offer useful details about the resources necessary to produce cutting-edge synthetic media with ML.

How to Build a Deepfake

Deepfakes are one specific application of ML, a field focused on the development of algorithms that improve as they process data. This processing results in a trained "model," a piece of software that ideally accomplishes the desired task. Consider the creation of an ML-driven face recognition system. The first step is to bring together a training dataset of both tagged photographs of faces and photographs containing no faces. During training, the ML algorithm learns from the provided examples to associate the images containing a face with the tag "face" and images without faces with the tag "no face." If done properly, the resulting trained model can process previously unseen images to determine if they contain a face.

This model gains a limited "understanding" of what a face looks like through the training process. This level of "understanding" is referred to in the field as a *representation*. Representations are at the core of how ML creates synthetic media. Specifically, engineers create faked media using a generative model—a class of models that can produce novel data similar to that used to train the system in the first place. A generative model trained on images of faces will gain a representation of what a face looks like. This representation can then produce new images of faces that have never existed.

"Deep learning"—from which "deepfake" draws its name—refers to a class of models used in ML known as neural networks; in recent years, they have proven to be some of the most successful means of constructing models. Researchers have taken special interest in generative models using deep learning—sometimes referred to as "deep generative models"—because of their proficiency at extracting representations from different kinds of data. This ability allows them to produce strikingly good imitations of their original training data.

The imitations produced by deep generative models are the "deepfakes" sparking public concern. This is an extremely active area of research: numerous models have been proposed in recent years that adopt different approaches with varying strengths and weaknesses. Some of the most prominent examples focus on the generation of images, including Glow (2018), PixelCNN (2016), NADE (2016), and DRAW (2015).[19]

While the research community has focused on static imagery, deep generative models have also succeeded in creating synthetic audio and video. Examples include WaveNet, a deep generative model released in 2016 that is adept at modeling audio data. WaveNet produces synthetic voices that sound significantly more natural and realistic than previous approaches.[20] Research in generative models has also identified several methods for producing synthetic video.[21]

One technique—a major source of the "deepfakes" most widely circulated beyond the research community—is known as generative adversarial networks, or GANs.[22] Invented in 2014, GANs have become a key area of exploration within the technical community. A variant of the basic ML workflow described above, GANs use two paired models. One model, the discriminative network, is trained on a dataset of interest, such as the corpus of tagged faces previously discussed. The second is a generative network, which produces synthetic data based on a given dataset.

A GAN places a generative network and a discriminative network into competition with one another.[23] The generator attempts to create novel data to "fool" the discriminator into classifying synthetic data erroneously as genuine. The game is played iteratively, and the successes and failures of the discriminator are used to train both networks in subsequent rounds.

As a result, the generator improves at creating imitations, and the discriminator improves its ability to detect fakes produced by the generator. If properly trained, the GAN results in a generator capable of creating fakes that closely resemble the original training data. Therefore, a GAN with a discriminator trained on images of faces would produce a generative network that can create novel, synthetic images of faces.

As with the other deep generative models described, researchers have illustrated the capacity of GANs to create realistic, synthetic imagery of objects in a range of contexts.[24] GANs have also demonstrated an impressive capacity to make seamless modifications in existing media. This includes the transformation of photos from day to night,[25] the artificial aging of faces,[26] and the substitution of animals with other animals in the same pose.[27]

Costs and Capabilities

Examining the technical literature on generative models helps determine the resources required to produce a high-quality deepfake, and the range of different kinds of faked media that can be generated. These facts hint at what kinds of malicious actors might use deepfakes for spreading disinformation, and the media they might create with the technology.

The research literature on synthetic face generation offers a good starting point for examining the costs and capabilities of generative models. With models trained

on faces, malicious actors might seek to produce believable profile photos for fake accounts on social media platforms or to create a false narrative around a made-up individual. One widely-cited paper from 2017 illustrates that state-of-the-art GANs can produce realistic, synthetic face images up to a 1024 x 1024 pixel resolution.[28] In this case, researchers relied on a training dataset of 30,000 photos of celebrities at the same resolution, and needed four days of training on eight Graphics Processing Units released in 2017.[29] GPUs—a specialized type of hardware—are a standard platform for conducting ML applications. Malicious actors willing to settle for generating lower-quality synthetic faces could do so with significantly smaller training datasets. GANs with smaller datasets—in the 2,000-image range—have generated synthetic images at a similar size but with less realism and a narrower range of faces.[30]

GANs trained solely on photographs of faces will be limited to synthetic face imagery, while disinformation perpetrators may desire access to models capable of simulating many kinds of objects. Recent work provides a few helpful datapoints for evaluating the resources needed to produce these more flexible models. SAGAN and BigGAN are two recent examples with state-of-the-art performance, both trained on ImageNet—a popular dataset consisting of more than 14 million images indexed into more than 20,000 categories.[31] SAGAN produces realistic synthetic imagery of a variety of objects at a 128 x 128 resolution after two weeks of training on a cluster of four GPUs.[32] BigGAN, which improves on the SAGAN performance at the 128 x 128 scale, produces larger, high-quality 512 x 512 images on 24 to 48 hours of training.[33] The scale and choice of hardware makes a significant difference in the time necessary to complete the training process. BigGAN was trained on 128 to 512 cores of a Google TPUv3 Pod, a hardware system released in 2018 specifically designed for accelerating machine learning applications.[34] The result is that disinformation campaigns with more money to spend on specialized hardware may create these kinds of flexible image generation models more quickly.

Disinformation actors will want to generate more than static images. A few examples provide a rough benchmark for the resources needed to apply deep generative models to other forms of media. WaveNet, the high-performance synthetic voice system previously discussed, was trained on 24.6 hours of English speech data, with an alternative demonstration trained on 34.8 hours of Chinese speech.[35] SampleRNN, an alternative speech synthesis model producing quality results, relied on a dataset of 20.5 hours of speech, training for approximately a week on a single GeForce GTX TITAN X, a GPU released in 2016.[36] As is the case with image-based generative models, campaigns with resources to invest in specialized hardware will be able to reduce training times for audio-based models. These training periods can be significantly shorter as a result of algorithmic improvements and differing

hardware setups. One recent paper on deep generative models for synthetic voice reported training times of 37.5 hours, leveraging a cluster of eight TITAN X GPUs with around 20 hours of speech data for training.[37]

Disinformation campaigns may also want to use generative models to create faked videos. The data, hardware, and training time needed to create high-performance models for video is generally greater than for images and audio. The technology is also less mature, leading to lower-quality fakes. One landmark 2018 demonstration of a model called "vid2vid" showed that GANs can successfully conduct what is known in the ML field as "video-to-video synthesis."[38] This process enables researchers to "translate" the visuals from one video into another: vid2vid can substitute objects, change video styling, and simulate entire scenes. The production of high-resolution 2K synthetic video of up to 30 seconds was achieved with an eight GPU cluster running over an approximately 10 day training process.[39] Researchers ran a number of experiments, with training datasets ranging from a collection of about 3,000 short street-scene videos to a 900-video collection of reporters providing briefings.[40] While providing a flexible framework for producing many kinds of synthetic video, the researchers note that the model continues to suffer from issues including the inability to "guarantee that an object has a consistent appearance across the whole video."[41]

Malicious actors may not need the ability to generate many different kinds of faked scenes offered by a model like vid2vid. A disinformation campaign may only want to leverage ML to fake a particular kind of video, such as a hoaxed recording of a political leader speaking at a podium. This would allow them to take advantage of narrower models developed by the ML field.

One prototypical example is "do as I do" video synthesis, which draws from a source video of a moving person and produces a synthetic video of a "target" individual in a separate scene doing the same motions. GANs have proven effective at producing these synthetic videos.[42] For a disinformation campaign, these narrower models are attractive because they require significantly less data and training time to create. One method developed by researchers in 2018 can accomplish "do as I do" synthesis effectively with far less data, using only 100 "source" videos of professional dancers for training, and five videos of "targets" making a range of motions for eight to 17 minutes.[43] More recent work on a system known as MetaPix shows that alternative approaches can achieve "do as I do" synthetic video production with only a few frames of data available from the "target."[44] This training for MetaPix was done on a cluster of four TITAN X GPUs, requiring a single day of training. However, quality problems persist. As with vid2vid, these generative models can noticeably distort the resulting video and struggle to synthesize believable motion of objects, such as loose clothing or hair.[45]

This review of the research literature understates the operational complexities faced by an online influence operation in deploying deepfakes "in the field." GANs are relatively delicate tools, even for trained researchers. Training is often unstable and subject to "mode collapse," in which the generative model arrives at a single or small set of outputs able to fool the discriminator, resulting in generators capable of producing only a tiny set of synthesized outputs.[46] These "collapsed" models will be useless to a disinformation campaign that needs to generate more than a handful of faked faces or images. The on-staff technical expertise of an influence operation will make a major difference in whether or not a malicious actor can create generative models effectively.

A disinformation campaign unwilling to deal with the cost and complexity of creating a deepfake from scratch could obtain a pre-trained model created by someone else. Increasingly, pre-trained models are being open-sourced or embedded in software for use by laypeople. Typically, these pre-trained deep generative models perform simple, routine transformations of a piece of media; this trend is most prominent in the "face swap," or the use of ML to replace one face in an image or video with another. FakeApp, an app released in 2018, enabled non-technical users to perform this transformation, frequently for the purpose of transplanting celebrity faces into pornographic films.[47] Today, the basic technology for creating fake swaps is now freely available in open-source software repositories online.[48] Freely or cheaply available generative models for creating a range of different fakes will likely become the norm as the knowledge to create deepfakes grows more widespread. While these pre-trained models may be lower quality and less customizable than models created from scratch, the low cost to using them may attract less well-resourced disinformation campaigns.

DEEPFAKE DETECTION

Online influence campaigns will not make the decision to use deepfakes in a vacuum. Beyond the resources needed to produce a deepfake and the resulting quality, malicious actors will weigh the risk of their hoaxed media being identified as fake.

To that end, progress in the field of ML on detecting deepfakes will influence how the technology is used to spread false narratives. Disinformation campaigns will avoid easily detectable deepfakes in favor of ones harder to identify. This section examines the current strengths and weaknesses in deepfake detection, and how detection algorithms might be used in practice.

Deepfake Detection: Strengths and Weaknesses

Digital media forensics—the field of research examining how faked or tampered media might be detected—has yielded a wealth of tools for identifying suspicious

artifacts out of sync with the known behaviors in cameras capturing images and video, as well as artifacts produced by post-capture processing and manipulation.[49] This field has begun to turn its attention to identifying synthetic media generated by GANs and deep generative models as they have grown in prominence in recent years.

One approach attempts to isolate and develop detection systems for artifacts and inconsistencies known to be common among different deepfake creation methods. GANs and other deep generative models produce faces exhibiting unrealistic behavior, such as infrequent blinking or a lack of subtle variations in skin color produced by heartbeats.[50] Other methods for identifying inconsistencies look for visual warping of certain elements like the face in synthetic media and the unusual coloration occasionally produced by GANs in images.[51]

ML has also been applied to the problem of detecting deepfakes.[52] Rather than looking for a specific inconsistency, ML-driven approaches focus on training models to examine many different features of a piece of media to identify a fake.

Researchers have reported strong success in identifying deepfakes with ML. One recent paper on a deep learning system called XceptionNet has claimed a 99.3 percent deepfake detection accuracy on raw images, and 81.0 percent accuracy on a more challenging task involving low-quality images. This model was trained on a dataset of more than 1.8 million manipulated images constructed from still frames of 1,000 deepfake videos.[53] An alternative system based on simpler ML methods has shown comparable performance with a smaller dataset of 225,000 images drawn from still imagery of faces.[54] These simpler methods also have the advantage of being trained on commodity CPUs, rather than requiring the more specialized GPU hardware needed to accelerate the training of large deep learning models.[55]

These ML-based detection algorithms suffer from one important drawback: they perform poorly when encountering novel means of creating faked media not incorporated into the original training set. Even beyond deepfake detection, ML models frequently perform well only on the dataset they were trained on, resulting in systems that fail when presented with new data.

Generalizing to new data has been a chronic problem with many of the deepfake detectors proposed in recent years. One evaluation of XceptionNet tested it against random deepfake videos from YouTube and concluded that its accuracy was "much, much lower" than reported. The training data used did not accurately represent the kinds of video manipulation being seen "in the wild."[56] Another 2018 paper evaluating multiple approaches noted the "deficiencies of detection algorithms when unknown data is presented" and concluded that "[these methods] cannot meet the challenge of detecting fake faces."[57]

Another concern is that many ML-based detection methodologies are vulnerable to "counter-forensics," or techniques a hypothetical hacker implements to subvert detection systems.[58] Deep generative models can be improved to avoid the suspicious distortions that indicate a fake image or video.[59] Subtle modifications to deepfakes can also camouflage them from popular ML-based detection methods.[60] These counter-forensic techniques will complicate the efforts of forensic experts in ascertaining media manipulation going forward.

Deepfake detection is far from perfect. While detection systems succeed in identifying known deepfake creation methods, they may generalize poorly to methods that are not incorporated into the training data. As a result, online influence operations have an opportunity to create and circulate faked media without detection.

Will Detection Be Deployed at Scale?

Success by researchers in detecting deepfakes is only one part of the equation. Deepfake detection systems are not likely to deter disinformation campaigns from spreading false narratives if the systems prototyped in the lab are never put to broader use by social media platforms, journalists, and others.

The integration of academic research into practical tools is uneven. Advances in detection are—by and large—not being integrated into standalone products or services for the general public. Only a few startup companies offer detection services commercially, but focus on servicing businesses concerned about the rise of deepfakes. Deeptrace, which claims to be the "first-to-market deepfakes detection solution," has followed this model, providing a proprietary detection technology for identifying synthetic media.[61]

Ubiquitous deepfake detection may not in itself deter malicious actors. Evidence suggests that the mere identification of faked content does not necessarily change public views about its veracity.[62] In this respect, recent commitments by major online platforms to remove deepfakes entirely may have more significant implications for online influence campaigns.

Public and policymaker concern has led to many of these platforms declaring policies against uploading deepfake content under certain circumstances. In January 2020, Facebook announced that it would begin removing content that "has been edited or synthesized...in ways that aren't apparent to an average person and would likely mislead someone into thinking that a subject of the video said words that they did not actually say" and were generated using AI or ML.[63] Twitter adopted a broader approach in February 2020, announcing it would remove and warn users against "synthetic or manipulated media that are likely to cause harm."[64] Platforms such as Reddit have adopted similar policies.[65]

Given the size of these platforms, effectively enforcing these policies will require internal systems to automatically detect when deepfakes have been uploaded. Major technology companies have thus accelerated research in detection. Since 2018, Google has made datasets of synthetic speech and video available to researchers in an effort to encourage the creation of common standards in the community and enable work on detection.[66] In 2019, Amazon Web Services, Facebook, Microsoft, and the Partnership on AI—a civil society organization—launched a "Deepfake Detection Challenge," a competition to encourage research and new approaches in identifying manipulated media.[67] These efforts will likely play a significant role both in pushing research forward and smoothing the progression of technical advance-ments into practical software solutions that can be implemented by intermediaries across the web.

The aggressive stance taken by online platforms against deepfakes will have implications for online influence operations. Malicious actors will face a more hostile environment for distributing deepfakes through major social media channels, as the content will be removed if identified by a detection algorithm. To the extent that disinformation campaigns create and spread deepfakes, they will do so when they can evade detection algorithms on these online platforms.

3 Future Scenarios and Recommendations

The research trends in deepfake creation and detection influence the resource requirements, media creation capabilities, and exposure risks that disinformation actors will face in spreading false narratives with synthetic media. These factors suggest two scenarios: one where disinformation campaigns adopt off-the-shelf ML models to "flood the zone" with deepfakes, and another where engineers create their own customized generative model to produce deepfakes for a targeted purpose. Of these two scenarios, "tailored" deepfakes are poised to be the greater threat over time.

This section of the paper assesses these two scenarios, tying together the disparate research trends discussed in the previous section to anticipate who is likely to use deepfakes in manipulating the media, how they will use them, and the overall threat posed by the technology. Based on this analysis, this section offers a set of recommendations to mitigate the threat.

COMMODIFIED DEEPFAKES: HEADED TO A STALEMATE

In one scenario, deepfakes could proliferate and become ubiquitous in the public discourse. Even poorly resourced disinformation perpetrators could take advantage of the technology as it becomes increasingly cheap and accessible. While deepfakes will become easier to produce, they will not pose a significant and growing threat. Quite the opposite: technical trends seem to suggest that detection systems will hold deepfakes at bay, confining them to limited areas of the web.

Deepfakes are indeed commodifying. The availability of open-source packages and off-the-shelf software succeeding face-swapping apps like FakeApp are likely just the beginning of a wave of tools for simple manip-

ulations of media. While ML models can be challenging to design from the ground up and train to a high level of performance, once built they are relatively easy to use and distribute. ML-driven media manipulation capabilities will become more—rather than less—accessible with time.

The spread of deepfake creation technology is not confined to underground or nefarious means: companies like Adobe have begun incorporating ML-enhanced tools for editing images and audio into their software.[68] A strong commercial motivation exists for platforms to offer these cutting-edge features to their customers.

The commercialization of deepfake creation tools and the open-sourcing of ML models will likely lower the cost and complexity of these tools to the point where even technically unsophisticated disinformation campaigns will adopt deepfakes. Where a simple face swap in an image or video might help the campaign harass their targets or spread uncertainty with little operational cost, it should be expected that they will do so.

But this increasing ubiquity overstates the actual impact. Current research shows that routine image and video modifications appear to be the easiest to detect. Moreover, widely accessible deepfake creation tools will likely be available to researchers and malicious actors alike, allowing researchers to examine the technology and develop detection schemes.

Even if researchers cannot access the underlying deepfake creation tool, the widespread use of a model for generating faked media produces ample training data for detection systems. Good examples include software like Face2Face and DeepFake—among the most widely used and most reported-on techniques for creating deepfakes. Researchers have focused their efforts on the synthetic media generated using these tools, making these techniques highly vulnerable to detection. As a method of producing deepfakes becomes more popular, detection teams will have access to the resources necessary to identify them.

Deepfakes created through "commodified" methods can therefore in theory be quickly identified. But effective detection will require continuous maintenance. The existing body of deepfake detection research suggests no conclusive solution for identifying ML-generated fakes will be found in the near future. Instead, detection systems will need constant updating on a case-by-case basis as new apps and pre-trained models emerge. Private companies are likely to take the lead on funding this upkeep. Nearly all major social media platforms have policies that ban malicious uses of deepfakes and face strong political pressures to combat disinformation.

Effective deepfake detection empowers companies to enforce policies on removing deepfakes from their platforms. The implementation of detection systems at the scale of platforms like Facebook and Twitter will make it difficult for commodified manipulations to spread broadly on these platforms before being taken down.

The malicious use of simple, commodity deepfakes will likely expand in the near term, but decline as major social media platforms implement the findings presently in the academic literature.

This is not to imply commodified deepfakes will have no impact on public discourse. Rather, these deepfakes will likely continue to circulate within specific, confined places online. Widespread, aggressive deepfake detection and takedown by the biggest online platforms will not cover private messaging networks, such as WhatsApp or Telegram, or smaller platforms such as Gab that have refused to filter content.[69] Some platforms may also opt to use detection systems to identify deepfakes for users rather than deleting the content, a permissive approach allowing these fakes to continue shaping perceptions.[70]

Disinformation campaigns will be able to influence specific audiences with commodified deepfakes on certain platforms rather than targeting public discourse at large. In these cases, the effectiveness of a deepfake will depend on the ability for individual users and small communities to weed out synthetic media as it appears. Whether deepfakes are able to influence discourse in this context will vary considerably from platform to platform and even among communities within a given platform.

TAILORED DEEPFAKES: A PERSISTENT THREAT

A disinformation campaign might choose to invest in constructing a custom ML model to produce deepfakes for a highly specific objective rather than relying on pre-existing tools and models. This could include the creation of a high-quality simulation of a target individual's voice, or the production of media integrating counter-forensics to defeat cutting-edge detection methods.

Unlike the commodified deepfakes discussed above, tailored deepfakes will pose a persistent if not growing threat over time. Disinformation campaigns can customize deepfakes to bolster specific false narratives a target audience may be susceptible to believing. These deepfakes will also be generated by custom ML models, enabling them to more effectively evade detection algorithms.

However, deep generative models will not serve malicious actors in all situations. Four important limitations constrain disinformation campaigns in effectively using the technology, offering predictions around who will use tailored deepfakes, when and how they will use them, and what they will attempt to fake using ML.

Who Can Build Tailored Deepfakes?

A review of the technical literature indicates that producing tailored deep generative models for malicious purposes is well within the reach of many, but not all, actors.

The specialized expertise needed to develop, train, and deploy custom ML systems is rapidly expanding, along with the body of freely available training mate-

rials.[71] Research literature on ML-generated synthetic media is openly available and described in sufficient detail to allow a person versed in the field to adapt them. This will make it easier over time for a malicious actor to obtain the knowledge or talent necessary for producing a deep generative model.

The specialized hardware needed to train a custom generative model is not prohibitively expensive. Many of the papers demonstrating state-of-the-art synthetic media creation require only a relatively small number of GPUs. The existing research suggests that a malicious actor could build a tailored GAN for producing synthetic faces with single digit numbers of high-end GPUs. At the time of writing, this hardware is available for purchase by the general public and generally costs in the range of $3,000 per GPU— an expense in the range of tens of thousands of dollars for a disinformation perpetrator.[72]

Disinformation actors may not even need to acquire and install their own specialized hardware. Cloud computing providers now rent access to GPUs for a fraction of the cost of outright purchase. Google currently offers access to the NVIDIA Tesla V100—an industry standard processor for ML applications—at $2.48 per hour per GPU.[74] Disinformation actors will increase their risks of exposure in using this infrastructure: most cloud providers collect identifying user information, and a record will exist of their activities that could be accessed by law enforcement or other government agencies. However, the opportunity to reduce the costs of creating deepfakes by an order of magnitude or more may outweigh these concerns.

These trends in talent and hardware suggest that tailored deepfakes are likely out of reach for an amateur troll, but quite affordable for a state, criminal organization, or political operation. Generating tailored deepfakes is somewhat more expensive than using existing media editing software to create faked media, but only marginally so for a well-resourced actor. The increased realism made possible through ML may well make it the preferred option.

When—The Disinformation Zero-Day

Sophisticated disinformation campaigns want deepfakes that will evade automated detection and takedown systems for as long as possible to maximize the audience viewing their content. Since existing detection methodologies for deepfakes appear strongest when dealing with a known method, influence operations will want to hold a custom deep generative model in reserve until a key moment: the week before an election, during a symbolically important event, or in a moment of great uncertainty.

Detection systems are vulnerable to what are known in the field of computer security as "zero-days." Zero-days are vulnerabilities unknown to defenders at the time an attacker deploys them. Often, little can be done to secure systems and

limit the damage when zero-days are first deployed, making them a major threat. A malicious actor able to develop a novel means of producing synthetic media "in house" could produce a kind of disinformation zero-day. These deepfakes will effectively evade exposure and takedown when released.

But, as is true in the computer security domain, the advantage an attacker reaps from using a zero-day is temporary. Once revealed, a vulnerability can be studied, and patches can be built that eliminate the effectiveness of the exploit. The same is true in synthetic media. Creating a vast corpus of fakes with a new deep generative model inadvertently produces training data for improving detection algorithms. These tailored models are therefore depleting assets: the more they are used, the less effective they are likely to become.

To this end, a disinformation campaign will not likely use a custom generative model as the "go to" form of content generation. Reserving tailored deepfakes for a key moment maximizes the impact of the fake, while limiting the possibility that detection systems will be trained to identify and weed out the content before it shapes public discourse.

How—The Problems of Training Time

Training a generative model to produce a high-quality deepfake takes time. ML is a computationally intensive procedure, often requiring a massive number of calculations in order to extract a good representation from the training data. This imposes significant operational constraints on disinformation campaigns, limiting how they can use tailored deepfakes to manipulate public discourse.

It is challenging to reliably make content "go viral." As a result, disinformation campaigns continuously improvise on existing narratives, seeking to ride emergent topics to prominence and appropriate them to their own advantage. This appears to be true even of the most well-resourced and large-scale campaigns of online propaganda. The 2016 Russian election interference campaign, for instance, produced a continuous stream of media that nimbly responded to trending memes and stories in the news cycle.[75] Many of these memes are momentary flashes in the pan—discussed only for a day or even a few hours before being replaced by another focus of discussion.

The rapidly fluctuating nature of online discourse suggests that deepfakes will not be an all-purpose content generation tool for disinformation campaigns. Training a relevant generative model will simply take too long.

The exact length of training time will depend on a number of factors, including the specific task the ML model is trained for, the computational power available, and even small details like the dimensions of the prospective image. But the training time to create a system to produce high-quality synthetic media can stretch from a

few days to a few weeks—and even this estimate may be optimistic. GANs are notoriously prone to failure in the training process, particularly if a disinformation campaign cannot recruit an experienced specialist in the field to assist in their operation.

These techniques can and will improve, but for now, the training time delay limits the usefulness of customized generative models for the flexible, rapid-fire content generation typical of many recent online influence efforts. It will be challenging to use deepfake technologies for unpredictable news events and viral "memes," as salient moments may lose relevance by the time a disinformation campaign aggregates data and trains a model for generating content. In these situations, a disinformation effort may find it faster and less expensive to use human agents to produce misleading content online.

The "sweet spot" for disinformation campaigns using customized deepfakes may lie elsewhere. First, malicious actors will likely use ML models to generate content for situations with sufficient lead time to prepare, train, and test their systems. Events that can be anticipated in advance or stories with ongoing narratives provide a stable set of circumstances for training a deep generative model. This can be more than a fixed date like an election: frequently recurring events such as mass shootings or protests provide evergreen "scenes" that a disinformation campaign might train a deep generative model to realistically simulate.

Second, disinformation campaigns will have incentives to invest in ML models that can be flexibly applied regardless of the specific situation at hand. For example, models that realistically simulate the voice of a high-value target could spread false narratives in a variety of contexts; the voice might be exploited to make a target appear as if they are opining on a range of different topics.

Third, ML models are likely to generate assets for a disinformation campaign when time pressure is not as acute. These are applications somewhat separate from the day-to-day content produced and distributed by the campaign. For instance, deep generative models may be used to improve the believability of a fake identity online by adding a high-quality profile photo. These elements do not need to be frequently updated, and a disinformation campaign may be able to populate this content with a generative model at greater leisure.

What—Data Accessibility Drives Content

Using tailored deepfakes in disinformation operations will require access to plentiful training data depicting the person, place or thing the campaign seeks to fake. Data is a critical input for creating ML models. The inability to compile a database of faces, for instance, prevents an actor from training a model to generate faces. Similarly, without voice samples, an ML system cannot simulate that voice realistically.

At the very least, sparse or difficult-to-access data for training will require a prospective disinformation actor to expend resources to acquire it. This may tip the

balance away from leveraging ML in certain cases where the costs of gathering data outweigh the benefits from generating more believable hoaxes. Conversely, available methods for acquiring datasets cheaply or freely will make ML use more likely.

The availability of data is therefore a useful lens to examine what malicious actors are more or less likely to depict in their deepfakes. While it is impossible to list the intersection of all places where a disinformation actor may seek to create synthetic media and where the data is plentiful, the research literature demonstrates a few relevant categories from a security perspective.

Public figures with a significant corpus of media available online are more susceptible to targeting in part because sufficient training data exists to imitate them. It is no surprise that researchers have tended to simulate major political figures and entertainment celebrities in papers on deep generative models: public recordings of them are plentiful and easy to access. In the video context, some public figures are more amenable to simulation than others because recordings depict them behaving in recurring patterns with small variation. There are many recordings of a politician speaking at a lectern, for instance, or a celebrity on the red carpet. This is an attractive target for disinformation efforts: the cost to acquire the data to train a generative model is low, and the individuals are high-profile. Similar logic applies to generating synthetic media depicting prominent landmarks or locations, where massive collections of photos or recordings may exist for training online.

On the other end of the spectrum, abundant data exists for commonplace objects and motions captured repeatedly in images and video. Images of faces, for instance, are easy to acquire, simplifying the creation of synthetic faces using ML. Videos that record common activities like dancing or singing are widely available, facilitating the simulation of new videos depicting those activities. These are also likely to be an attractive target for malicious actors: it may be useful to simulate events with anonymous crowds in locations that are hard to verify. Disinformation efforts have frequently relabeled and misleadingly contextualized images and video from unrelated situations to drive false narratives.[76] They might do the same with a scene generated from scratch.

Where is data scarce? What are the circumstances under which a malicious actor might have difficulty training a tailored deep generative model? For now, data is scarce when a malicious actor seeks to depict a *specific* individual, object, or scene that has not been widely recorded. Creating customized synthetic media around an individual who becomes an unexpected internet celebrity due to a small number of "viral" videos may be difficult through ML. Similarly, a specific building or location previously largely ignored but under significant scrutiny as the site of a major news story may be difficult to simulate. In these cases, disinformation campaigns may be

delayed while they wait for sufficient training data to accrue online before they can train a generative model or be forced to spread their narrative through alternative means.

RECOMMENDATIONS

Deepfake creation will become easier with time. The material costs are not prohibitively high, and much of the know-how is published openly and available to use. Costs will continue to decline as research advances are distilled into easy-to-use software and open source code. However, detection systems and takedown policies will help to keep commodified deepfake technologies somewhat at bay. The greater threat will be from tailored, targeted fakes: the trends indicate that these hoaxes will remain a persistent risk.

This analysis offers concrete recommendations for confronting these threats and limiting the impact of deepfakes in manipulating public discourse. These recommendations seek to hamper both the influence of commodified deepfakes and the use of the technology by sophisticated malicious actors creating generative models for their disinformation efforts.

Recommendation 1: Build a Deepfake "Zoo"

Absent some major technical breakthrough, deepfake detection will evolve as a cat-and-mouse game. Novel means of creating synthetic media will be invented, and detection systems trained to account for the new method.

Success will therefore depend on how quickly detection systems used by giant social media platforms and smaller entities can account for new methods. Rapid integration means disinformation campaigns will confront a hostile environment where synthetic media is quickly identified and removed before proliferating. If the time between the first use of a new technique and its widespread integration into detection systems is sufficiently narrowed, it may render the use of ML for these purposes an unattractive option for malicious actors.

Making this integration work will require rapid access to samples of media produced by different deepfake models. This is particularly important when the threat emerges from a sophisticated attacker able to construct novel generative systems to augment a disinformation campaign. In these cases, media examples produced by a novel technique may be initially sparse. It will be important to identify and aggregate examples for training data as quickly as possible.

In order to accelerate this process, stakeholders—platforms, researchers, companies—should invest in the creation of a deepfake "zoo" that continuously aggregates and makes freely available datasets of synthetic media as they appear online. These datasets would be categorized by media type, and if possible, include annotations about the likely method used to create the content. This would mirror

initiatives taken by organizations like the National Cyber-Forensics and Training Alliance in the cybersecurity domain. The NCFTA operates as a common clearinghouse for data about new and emerging malware threats, enabling their partners in private industry and government to respond more effectively.[77]

Similarly, by lowering the costs of acquiring relevant, up-to-date training data to augment detection algorithms, the "zoo" would make overall detection more robust. This would improve on the infrequently updated set of common datasets in use in the research community.

Recommendation 2: Encourage Better Capabilities Tracking

Inconsistent practices for documenting research have already led members of the technical community to raise concerns about a looming reproducibility crisis in ML.[78] Beyond posing problems for the verification and validation of research results, inconsistent documentation makes precise evaluations of the availability of ML technologies for various applications difficult. This is clear in the review of the technical literature around deep generative models. Papers are inconsistent in their disclosures of key facts such as the hardware used in the training process, characteristics of the training data, and length of the training process.[79] As a result, this paper must rely on a smaller set of research that has documented these aspects of the experimental process.

Inconsistent documentation poses a significant issue in assessing the current state and future prospects of media manipulation and deep generative models. It is difficult to ascertain the speed at which research advances make it possible for certain actors to produce cutting-edge synthetic media at a low cost, hindering threat assessment and the effective allocation of resources.

Research communities, funding organizations, and academic publishers should work toward developing common standards for reporting progress in generative models. This might include raising the bar on documenting the processes used in training a new model, as well as integrating this information in a machine-readable way into the metadata included with published academic papers. Such standardization would improve transparency around the state of the field in ways that facilitate better strategic planning.

Recommendation 3: Commodify Detection

Simple deepfakes can still pose a threat in a world where detection systems are widely implemented. While these fakes will be quickly detected and removed on the most popular, mainstream platforms for distributing content, they will still spread in the less monitored spaces of the web. This includes distribution through private messaging platforms, which already serve as channels for false narra-

tives even in the absence of ML-generated fakes.[80] This content will also continue to spread among smaller platforms with a more hands-off approach to synthetic content.

In these cases, the spread of a deepfake depends on the receptiveness of the viewer, rather than the effectiveness of a detection algorithm. The research literature around the sociological and psychological forces that drive individuals to believe or spread disinformation is expansive and contains many open questions.[81] In some cases, a fact check tool or deepfake identification algorithm will do little to dissuade someone from accepting a false narrative. Corrections may even backfire and increase misperceptions in certain individuals.[82]

Regardless, citizens should have access to the tools and training that enable them to investigate a suspicious piece of media if they so choose. A trained eye can identify crude deepfakes without any special procedures and processes. It may be important in this context to raise public awareness about deepfakes and to highlight indicative examples. Regular training sessions for journalists and people in professions likely to be targeted may also help limit the extent to which members of the public are duped.[83]

In parallel, philanthropic organizations and government agencies should give grants that facilitate the translation of research findings in deepfake detection into user-friendly apps for analyzing media that members of the public might encounter while browsing the web. This investment would strengthen the passive resistance that the public has against synthetic media generated by disinformation campaigns. It would also improve situational awareness for those seeking to mitigate the threat: these apps might aid in early identification of new techniques for producing deepfakes long before the research community or the mainstream media would otherwise be aware of it.

Recommendation 4: Proliferate "Radioactive" Data

Detection and attribution are major concerns for the disinformation actor. When quickly identified and taken down, deepfakes cannot influence public discourse. Even worse for the perpetrator, investigators may be able to uncover previously concealed elements of a disinformation campaign when synthetic media contains distinctive traces. These indicators could include the false identities used to distribute content and inferences about the target and objectives of the campaign.

At the same time, disinformation actors will leverage ML only to the extent that it is cost-effective. They will frequently rely on free, publicly available images and video for training to avoid assuming the heavy costs of acquiring and structuring data. This open-source reliance creates a vulnerability that could raise the risks for disinformation actors seeking to leverage ML for malicious purposes.

ML researchers have recently demonstrated a method that enables datasets to be made "radioactive," containing traces non-obvious to the human eye, but later extractable from media produced by models trained on that data.[84] Usage of "radioactive" data can be detected even when it constitutes as little as one percent of the data used to train the model.[85] These subtle modifications do not significantly affect the performance of models trained on marked datasets.

Deepfakes trained on "radioactive" data can be easily identified, offering a way to check whether an image or video is synthetically generated by an ML model without elaborate media forensics techniques. The unwary disinformation actor might draw on publicly available data, train a generative model, and produce synthetic media all without knowing that their training corpus has been marked. Even if the tainted dataset is combined with others prior to training, these markers would persist in the resulting media.

Stakeholders interested in mitigating the harm from deepfakes should encourage the "radioactive" marking of public datasets likely to be used as raw material for training deep generative models. For instance, prior to public release, academic and corporate ML labs might mark datasets of faces, voices, and other data of high value to disinformation efforts. These alterations would not materially harm research and legitimate applications, but would allow for quick detection of synthetic media circulating online that is purported to be legitimate. This would significantly lower the costs of detection for commodified deepfakes produced by pre-trained models and used by less technically sophisticated actors.

Widespread implementation of a variety of marking techniques would raise the risks and costs for sophisticated disinformation actors, as well. These campaigns would either be forced to source their own datasets to ensure the media they produce was unmarked, or implement processes for "cleaning" radioactive datasets. Marking datasets in this manner would raise the level of uncertainty for disinformation campaigns. Faced with the possibility that their datasets might contain hidden radioactivity, disinformation actors might opt out of the technology altogether.

Conclusion

A dramatic demonstration in the lab often reveals little about how a technology will be used in the real world. Deepfakes are no exception. While the use of ML to produce sharp, high-fidelity synthetic media is an impressive technical feat, the incentives of malicious actors will shape the ultimate threat the technology poses. Policymakers and national security researchers should avoid giving in to hype, but rather take precautions when sensible.

Deepfakes are not magic: ML is not yet so advanced that it can effortlessly conjure up fake scenes indistinguishable from reality. There is a real cost in using ML. Training data, computational power, and technical expertise must all be assembled to use it effectively. Limitations in the methodology constrain what fakes can be made, and how quickly they can be generated. Moreover, constantly evolving detection methods can make synthetic media easier to identify "in the wild."

These real, somewhat humdrum considerations provide crucial hints toward how a disinformation campaign is likely to use this technology to manipulate public discourse. While commodification will make deepfakes ever easier to produce, off-the-shelf technology for producing synthetic media will also become easier to detect and filter automatically. This limits the impact of this technology on mainstream platforms and narrows their scope to less monitored areas of the web.

The greater threat is likely from a sophisticated disinformation effort that tailors ML models for particular purposes. Moderately well-resourced disinformation efforts can afford custom generative models that produce cutting-edge deepfakes, but even in these cases, malicious actors are

constrained. The strategic dynamics of detection, the demands of training time, and accessibility of data all conspire to make some operational uses of deepfakes likelier than not.

This review also suggests a set of high-impact interventions helping to limit the effect of deepfakes in the media manipulation space. These recommendations include developing a "zoo" of samples produced by various deep generative models, better documentation of generative models within the research literature, funding to support the commodification of detection tools, and the use of "radioactive" data to strike at malicious actors who would use public training sets for harmful purposes.

While deepfakes have received outsized attention from policymakers and the popular press, less visually striking tactics may exert a greater influence on online disinformation campaigns over time. Conversational AI systems might be used to make large swarms of fake identities more believable and persuasive. Predictive algorithms could enable malicious actors to better and more subtly target individuals and communities receptive to their messaging. A similar approach grounded in the technical literature should be taken to examine these capabilities, as well.

Finally, strategic thinking about the intersection of ML and disinformation should not lose sight of the people behind the screen. Disinformation spreads in part because fake content appears genuine on its face and resists forensic attempts to identify tampering or manipulation. Equally important is the beholder, who must believe the false narrative depicted in a fake and share it widely. This analysis can spotlight places where disinformation campaigns might leverage modern technology to strike, but their ultimate success will depend on the receptiveness of their audiences. Even as ML continues to rapidly advance, these social and psychological dimensions of influence and disinformation will remain critical to the construction of an effective defense in this domain.

Glossary

- Artificial intelligence: The field of computer science research focused on enabling machines to have "intelligence," broadly defined.
- Deepfake: Colloquial term broadly referring to synthetic media produced using tools from the field of machine learning.
- Deep learning: A family of techniques in machine learning that relies on the use of "neural networks," specific method for constructing models.
- Generative adversarial network (GAN): One particular application of machine learning that results in models that can produce strikingly high-quality synthetic media. The core technology behind many of the prominent deepfakes that have circulated through the web in recent years.
- Machine learning: The subfield of artificial intelligence focused on the design of models that improve through processing of data, referred to in the field as "training."
- Model: A piece of software used in the machine learning training process that improves as it processes data. The trained model is then used to accomplish a desired task.
- Synthetic media: Media which is not authentically recorded from the real world, but instead faked using a variety of different techniques.

Endnotes

1. NOVA, "Deepfake Videos Are Getting Terrifyingly Real", YouTube, accessed March 4, 2020, https://www.youtube.com/watch?v=T76bK2t2r8g; Kevin Roose, "Here Come the Fake Videos, Too," *The New York Times*, March 8, 2018, https://www.nytimes.com/2018/03/04/technology/fake-videos-deepfakes.html.
2. Cara Curtis, "Deepfakes are being weaponized to silence women — but this woman is fighting back," *The Next Web*, October 8, 2018, https://thenextweb.com/code-word/2018/10/05/deepfakes-are-being-weaponized-to-silence-women-but-this-woman-is-fighting-back/.
3. Office of the Director of National Intelligence, *Background to 'Assessing Russian Activities and Intentions in Recent US Elections*, January 2017, https://www.dni.gov/files/documents/ICA_2017_01.pdf.
4. Regina Rini, "Deepfakes Are Coming. We Can No Longer Believe What We See," *New York Times*, June 10, 2019, https://www.nytimes.com/2019/06/10/opinion/deepfake-pelosi-video.html.
5. *National Security Challenges of Artificial Intelligence, Manipulated Media, and 'Deepfakes'*, 116th Cong. (2019) (statement of Adam B. Schiff, Chairman of the House Permanent Select Committee on Intelligence), https://docs.house.gov/meetings/IG/IG00/20190613/109620/HHRG-116-IG00-MState-S001150-20190613.pdf.
6. "The women missing from the silver screen and the technology used to find them," Google, accessed March 27, 2020, https://about.google/intl/en_us/main/gender-equality-films/ (gender bias); Jake Swearingen, "A.I. is Flying Drones (Very, Very Slowly)," *The New York Times*, March 26, 2019, https://www.nytimes.com/2019/03/26/technology/alphapilot-ai-drone-racing.html (drone piloting).
7. Office of the Director of National Intelligence, *Background*.
8. David Lee, "The tactics of a Russian troll farm," BBC, February 16, 2018, https://www.bbc.com/news/technology-43093390.
9. Alice Marwick and Rebecca Lewis, *Media Manipulation and Disinformation Online* (New York, NY: Data and Society Research Institute, 2017), https://datasociety.net/library/media-manipulation-and-disinfo-online/.
10. Rachel Lerman, "Video of Pelosi brings renewed attention to 'cheapfakes,'" *Associated Press*, February 10, 2020, https://apnews.com/12443c46b8cfee5e9659abb31eee5142; Lisa Fazio, "Who needs deepfakes? Simple out-of-context photos can be a powerfully low-tech form of misinformation," *Nieman Lab* (blog), *Nieman Foundation*, February 24, 2020, https://www.niemanlab.org/2020/02/who-needs-deepfakes-simple-out-of-context-photos-can-be-a-powerfully-low-tech-form-of-misinformation/.
11. Taylor Hatmaker, "What we can learn from the 3,500 Russian Facebook ads meant to stir up U.S. politics," *Techcrunch*, May 10, 2018, https://techcrunch.com/2018/05/10/russian-facebook-ads-house-intelligence-full-list/.

12. This has been recognized as an important characteristic of the Russian approach to these types of operations, see Molly Mckew, "The Gerasimov Doctrine," *Politico*, Sept/Oct 2017, https://www.politico.com/magazine/story/2017/09/05/gerasimov-doctrine-russia-foreign-policy-215538. Researchers Bobby Chesney and Danielle Citron have also explored how deepfakes may erode trust in legitimate media in "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security." Public Law Research Paper No. 692, University of Texas Law, Austin, TX, July 2018. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954.
13. Lerman, "Video of Pelosi."
14. Stuart Dredge, "Five of the best face swap apps," *The Guardian*, March 17, 2016, https://www.theguardian.com/technology/2016/mar/17/five-of-the-best-face-swap-apps.
15. Social media platforms have been increasingly aggressive in deplatforming influence campaigns. Nicolas Fandos and Kevin Roose, "Facebook Identifies an Active Political Influence Campaign Using Fake Accounts," *New York Times*, July 31, 2018, https://www.nytimes.com/2018/07/31/us/politics/facebook-political-campaign-midterms.html; Josh Taylor, "Twitter deletes 170,000 accounts linked to China influence campaign," *The Guardian*, June 12, 2020, https://www.theguardian.com/technology/2020/jun/12/twitter-deletes-170000-accounts-linked-to-china-influence-campaign.
16. Monica Bickert, "Enforcing Against Manipulated Media," *Facebook Newsroom*, January 6, 2020, https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/; Yoel Roth and Ashita Achuthan, "Building rules in public: Our approach to synthetic & manipulated media," *Twitter* (blog), February 4, 2020, https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html.
17. Shane Barratt and Rishi Sharma, "A Note on the Inception Score." Preprint, submitted January 6, 2018, https://arxiv.org/abs/1801.01973; Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahair, "How good is my GAN?" Preprint, submitted July 25, 2018, https://arxiv.org/abs/1807.09499.
18. Mario Lucic et al., "Are GANs Created Equal? A Large-Scale Study." Preprint, submitted November 28, 2017, http://arxiv.org/abs/1711.10337.
19. Diederik Kingma and Prafulla Dhariwal, "Glow: Generative Flow with Invertible 1x1 Convolutions." Preprint, submitted July 9, 2018. https://arxiv.org/abs/1807.03039 (Glow); Aaron van den Lord, Nal Kalchbrenner, and Koray Kavukcuoglu, "Pixel Recurrent Neural Networks." Preprint, submitted January 25, 2016, https://arxiv.org/abs/1601.06759 (PixelCNN); Benigno Uria et al., "Neural Autoregressive Distribution Estimation." Preprint, submitted May 7, 2016, https://arxiv.org/abs/1605.02226 (NADE); Karol Gregor et al., "DRAW: A Recurrent Neural Network for Image Generation." Preprint, submitted May 20, 2015, http://arxiv.org/abs/1502.04623 (DRAW).
20. Aaron van den Oord et al., "WaveNet: A Generative Model for Raw Audio." Preprint, submitted September 19, 2016, http://arxiv.org/abs/1609.03499.
21. Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba, "Generating Videos with Scene Dynamics." Preprint, submitted September 8, 2016, https://arxiv.org/abs/1609.02612.
22. For a non-technical introduction to GANs, see Cade Metz, "Google's Dueling Neural Networks Spar to Get Smarter, No Humans Required," Wired, April 11, 2017, https://www.wired.com/2017/04/googles-dueling-neural-networks-spar-get-smarter-no-humans-required/.
23. For more detail on this process, see Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning* (Cambridge: MIT Press, 2016), Chapter 20.10.4, https://www.deeplearningbook.org/contents/generative_models.html.

24. Augustus Odena, Christopher Olah, and Jonathon Shlens, "Conditional Image Synthesis With Auxiliary Classifier GANs." Preprint, submitted October 30, 2016, https://arxiv.org/abs/1610.09585; Marco Marchesi, "Megapixel Size Image Creation using Generative Adversarial Networks." Preprint, submitted March 31, 2017, https://arxiv.org/abs/1706.00082; David Berthelot, Thomas Schumm, and Luke Metz, "BEGAN: Boundary Equilibrium Generative Adversarial Networks." Preprint, March 31, 2017, https://arxiv.org/abs/1703.10717.
25. Ming-Yu Liu, Thomas Breuel, and Jan Kautz, "Unsupervised Image-to-Image Translation Networks." Preprint, submitted March 2, 2017, http://arxiv.org/abs/1703.00848.
26. Grigory Antipode, Moez Baccouche, and Jean-Luc Dugelay, "Face Aging With Conditional Generative Adversarial Networks." Preprint, submitted February 7, 2017, https://arxiv.org/abs/1702.01983.
27. Jun-Yan Zhu et al., "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks." Preprint, submitted March 30, 2017, http://arxiv.org/abs/1703.10593.
28. Tero Karras et al., "Progressive Growing of GANs for Improved Quality, Stability, and Variation." Preprint, submitted October 27, 2017, https://arxiv.org/abs/1710.10196; Tero Karras, Samuli Laine, and Timo Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks." Preprint, submitted December 12, 2018, https://arxiv.org/abs/1812.04948 (a more recent paper showing a similar training arrangement).
29. Karras et al., "Progressive Growing"; Karras et al., "A Style-Based Generator."
30. Marchesi, "Megapixel," 1.
31. "ImageNet," Stanford Vision Lab, accessed March 4, 2020, http://www.image-net.org/.
32. Han Zhang et al., "Self-Attention Generative Adversarial Networks." Preprint, May 21, 2018, https://arxiv.org/abs/1805.08318.
33. Andrew Brock, Jeff Donahue, and Karen Simonyan, "Large Scale GAN Training for High Fidelity Natural Image Synthesis." Preprint, submitted September 28, 2018, https://arxiv.org/abs/1809.11096.
34. Brock et al., "Large Scale GAN."
35. van den Oord et al., "Wavenet"; Yuxuan Wang et al., "Tacotron: Towards End-to-End Speech Synthesis." Preprint, submitted March 29, 2017, http://arxiv.org/abs/1703.10135.
36. Soroush Mehri, "SampleRNN: An Unconditional End-to-End Neural Audio Generation Model." Preprint, submitted February 11, 2017, http://arxiv.org/abs/1612.07837.
37. Sercan O. Arik et al., "Deep Voice: Real-time Neural Text-to-Speech." Preprint, submitted February 25, 2017, http://arxiv.org/abs/1702.07825.
38. Ting-Chun Wang, "Video-to-Video Synthesis." Preprint, submitted August 20, 2018, https://arxiv.org/abs/1808.06601.
39. Wang, "Video-to-Video."
40. Wang, "Video-to-Video."
41. Wang, "Video-to-Video."
42. For a brief review of recent literature on these techniques, see Caroline Chan et al., "Everybody Dance Now." Preprint, submitted August 22, 2018, http://arxiv.org/abs/1808.07371.
43. Chan et al., "Everybody."

44. Jessica Lee, Deva Ramanan and Rohit Girdhar, "MetaPix: Few-Shot Video Retargeting." Preprint, submitted October 10, 2019, http://arxiv.org/abs/1910.04742.

45. Chan et al., "Everybody."

46. "Common Problems", Google, accessed March 4, 2020, https://developers.google.com/machine-learning/gan/problems.

47. Samatha Cole, "AI-Assisted Fake Porn Is Here and We're All Fucked," *Motherboard*, December 11, 2017, https://www.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn.

48. "DeepFaceLab," Github, accessed March 4, 2020, https://github.com/iperov/DeepFaceLab.

49. Two useful surveys reviewing techniques in this field include Raahat Devender Singh and Naveen Aggarwal, "Video content authentication techniques: a comprehensive survey," *Multimedia Systems* 24 (2018): 221-240, https://link.springer.com/article/10.1007/s00530-017-0538-9 and Luisa Verdoliva, "Media Forensics and DeepFakes: an overview." Preprint, submitted January 18, 2020, https://arxiv.org/abs/2001.06564.

50. Yuezun Li, Ming-Ching Chang, and Siwei Lyu, "In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking." Preprint, submitted June 7, 2018, https://arxiv.org/abs/1806.02877.

51. Verdoliva, "Media Forensics," 10.

52. Verdoliva, "Media Forensics," 10-12.

53. Andreas Rossler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images." Preprint, submitted January 25, 2019, https://arxiv.org/abs/1901.08971.

54. Xin Yang et al., "Exposing GAN-synthesized Faces Using Landmark Locations." Preprint, submitted March 30, 2019, http://arxiv.org/abs/1904.00167; Yuezun Li and Siwei Lyu, "Exposing DeepFake Videos By Detecting Face Warping Artifacts." Preprint, submitted November 1, 2018, http://arxiv.org/abs/1811.00656.

55. Yang, "Exposing GAN-synthesized Faces".

56. Rayhane Mama and Sam She, "Towards Deepfake Detection That Actually Works," *Dessa* (blog), November 24, 2019, https://www.dessa.com/post/deepfake-detection-that-actually-works.

57. Ali Khodabakhsh et al., "Fake Face Detection Methods: Can They Be Generalized?" Paper presented at the 2018 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, September 2018, https://ieeexplore.ieee.org/document/8553251/.

58. Owen Mayer and Matthew Stamm, "Countering Anti-Forensics of Lateral Chromatic Aberration." Paper presented at IH&MMSec, Philadelphia, PA, June 2017, http://misl.ece.drexel.edu/wp-content/uploads/2017/08/Mayer_IHMMSec_2017.pdf.

59. Verdoliva, "Media Forensics," 15-16.

60. Verdoliva, "Media Forensics," 15-16.

61. "Deeptrace", Deeptrace Labs, accessed March 4, 2020, https://deeptracelabs.com/.

62. Gordon Pennycook et al., "The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings." Preprint, submitted September 14, 2017, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3035384.

63. Bickert, "Enforcing Against Manipulated Media."

64. Roth and Achuthan, "Building rules in public."

65. Jay Peters, "Reddit bans impersonation on its platform," *The Verge*, January 9, 2020, https://www.theverge.com/2020/1/9/21058803/reddit-account-ban-impersonation-policy-deepfakes-satire-rules.
66. Nick Dufour and Andrew Gully, "Contributing Data to Deepfake Detection Research," *Google AI* (blog), *Google*, September 24, 2019, https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html.
67. Claire Leibowicz, *A Report on the Deepfake Detection Challenge* (San Francisco, CA: The Partnership on AI, 2020), https://www.partnershiponai.org/a-report-on-the-deepfake-detection-challenge/.
68. "Adobe Sensei", Adobe, accessed March 4, 2020, https://www.adobe.com/sensei.html.
69. Emma Grey Ellis, "Gab, the Alt-Right's Very Own Twitter, Is The Ultimate Filter Bubble," *Wired*, September 14, 2016, https://www.wired.com/2016/09/gab-alt-rights-twitter-ultimate-filter-bubble/.
70. Pennycook, "The Implied Truth Effect."
71. For some metrics on the rate of change in educating engineers proficient in ML techniques, see Raymond Perrault, et al., *Artificial Intelligence Index Report 2019* (Stanford, CA: Human-Centered AI Institute, 2019), Chapter 5, https://hai.stanford.edu/sites/g/files/sbiybj10986/f/ai_index_2019_report.pdf.
72. "NVIDIA Titan RTX", NVIDIA, accessed March 4, 2020, https://www.nvidia.com/en-gb/deep-learning-ai/products/titan-rtx/.
73. "GPUs pricing", Google Cloud, accessed June 30, 2020, https://cloud.google.com/compute/gpus-pricing.
74. Sharad Goal, Duncan Watts, and Daniel Goldstein, "The structure of online diffusion networks," *Proceedings of the 13th ACM Conference on Electronic Commerce* (2012): 623-638, https://dl.acm.org/doi/10.1145/2229012.2229058.
75. Lee, "The tactics."
76. "Out-of-Context Photos Are a Powerful Low-Tech Form of Misinformation", *Snopes News* (blog), *Snopes*, February 15, 2020, https://www.snopes.com/news/2020/02/15/out-of-context-photos-are-a-powerful-low-tech-form-of-misinformation/.
77. "Malware and Cyber Threat Program", National Cyber-Forensics and Training Alliance, accessed June 29, 2020, https://www.ncfta.net/malware-and-cyber-threat-program/.
78. Gregory Barber, "Artificial Intelligence Confronts a 'Reproducibility' Crisis," *Wired*, September 16, 2019, https://www.wired.com/story/artificial-intelligence-confronts-reproducibility-crisis/.
79. Matthew Hutson, "Artificial intelligence faces reproducibility crisis," *Science*, February 16, 2018, 725-726.
80. Daniela Flamini, "WhatsApp efforts to curb misinformation aren't entirely effective, research shows," *Poynter*, September 27, 2019, https://www.poynter.org/fact-checking/2019/whatsapp-efforts-to-curb-misinformation-arent-entirely-effective-research-shows/; Sharon Moshavi, "Op-ed: Shining light into the dark spaces of chat apps," *Columbia Journalism Review*, January 14, 2020, https://www.cjr.org/opinion/whatsapp-messenger-misinformation.php.
81. For a sample of the vast literature on these topics, see Joshua Tucker et al., *Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature* (Menlo Park, CA:

Hewlett Foundation, 2018), https://www.hewlett.org/wp-content/uploads/2018/03/Social-Media-Political-Polarization-and-Political-Disinformation-Literature-Review.pdf.

82. Brendan Nyhan and Jason Reifler, "When Corrections Fail: The Persistence of Political Misperception," *Political Behavior* 32 (2010): 303-330, https://link.springer.com/article/10.1007/s11109-010-9112-2. The prevalence of the "backfire effect" has been challenged in subsequent studies, for example in Thomas Wood and Ethan Porter, "The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence," *Political Behavior* 41 (2019): 135-163, https://link.springer.com/article/10.1007/s11109-018-9443-y.
83. For an example of what this training might look like, see First Draft News, "First Draft launches its online verification training course," *First Draft* (blog), October 11, 2017, https://firstdraftnews.org/latest/course-launches/.
84. Alexandre Sablayrolles, Matthijs Douze, and Hervé Jégou, "Using 'radioactive data' to detect if a data set was used for training," *Facebook Artificial Intelligence* (blog), February 5, 2020, https://ai.facebook.com/blog/using-radioactive-data-to-detect-if-a-data-set-was-used-for-training/.
85. Sablayrolles et al., "Using 'radioactive data'."

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

CSET.GEORGETOWN.EDU | CSET@GEORGETOWN.EDU