978-9934-564-49-9

DEEPFAKES – PRIMER AND FORECAST

Published by the
NATO Strategic Communications
Centre of Excellence

ISBN: 978-9934-564-49-9
Author: Tim Hwang
Copy Editing: Jazlyn Malnczyk
Design: Kārlis Ulmanis

Riga, May 2020
NATO STRATCOM COE
11b Kalnciema Iela
Riga LV1048, Latvia
www.stratcomcoe.org
Facebook/stratcomcoe
Twitter: @stratcomcoe

Tim Hwang is Director of the Harvard-MIT Ethics and Governance of AI Initiative, a philanthropic project working to ensure that machine learning and autonomous technologies are researched, developed, and deployed in the public interest. Previously, he worked at Google, where he was the company's global public policy lead on artificial intelligence, leading outreach to government and civil society on issues surrounding the social impact of the technology. Dubbed "The Busiest Man on the Internet" by Forbes Magazine, his current research focuses on the geopolitical aspects of computational power and machine learning hardware, and the strategy of modern information warfare.

This paper provides a primer to deepfakes and forecasts their potential future role in online disinformation campaigns. It concludes that while the threat from deepfakes is real, the risks are somewhat narrower than is frequently portrayed. It also argues that deepfakes may serve as a distraction, reducing focus on the deeper issues that must be resolved to confront the problems of online disinformation and misinformation.

INTRODUCTION

"Artificial intelligence"—the broad category of study exploring the creation of intelligent machines—has enjoyed a resurgence in the last decade, driven by a combination of research breakthroughs, a massive expansion in access to data, and advances in computational hardware. New developments and applications have captured the imagination of policymakers and the public at large, inspiring both hopes and fears around artificial intelligence and its future prospects.

Among the many areas of concern around the technology, perhaps one of the most widely discussed has been the threat posed by "deepfakes": synthetic audio, images, and video generated with artificial intelligence. Deepfakes are often strikingly realistic and sometimes challenging to distinguish from the genuine article. Artificial intelligence has been used to produce deepfakes depicting prominent political figures from Donald Trump to Vladimir Putin saying a variety of things they never in fact said.[1]

The technology used to produce this faked media has far-reaching implications for art as well as science, but this paper focuses on its potential impact on disinformation, in particular, political propaganda and social manipulation. In an era in which disinformation campaigns and online media manipulation efforts are at the top of the

agenda, these developments appear poised to make these existing threats worse. As the headline for one *New York Times* op-ed put these concerns succinctly, "Deepfakes are Coming. We Can No Longer Believe What We See."[2]

There appears to be a range of situations where deepfakes could be used to harmful effect. Could a well-timed video hoax be launched by a malicious actor to disrupt an election? Might a massive wave of high-fidelity fakes swamp social media platforms and further erode trust in the media and institutions? Could deepfakes produce a "liar's dividend", allowing anyone to simply reject any evidence of their wrongdoing as a high-tech fake?[3]

This paper provides a primer to deepfakes and forecasts their potential future role in online disinformation campaigns. It concludes that while the threat from deepfakes is real, the risks are somewhat narrower than is frequently portrayed. It also argues that deepfakes may serve as a distraction, reducing focus on the deeper issues that must be resolved to confront the problems of online disinformation and misinformation.

This paper is divided into three parts. Part I explores recent developments in the technical field of artificial intelligence and how they have facilitated the creation of deepfakes. Part II will examine the major trends that are shaping the strategic landscape around deepfakes. Part III extrapolates from those trends to assess the future of deepfakes and offers some recommendations to those seeking to address the threat.

ARTIFICIAL INTELLIGENCE AND DEEPFAKES

"Artificial intelligence"—a broad and non-technical phrase—evokes the realm of science fiction. But the technology that creates deepfakes is neither fiction nor magic. To accurately assess the real threat posed by deepfakes, it is critical to examine precisely how they are produced. This enables a better understanding of where the technology may pose an actual risk, and where fears may exceed the real-life capabilities of these new technologies.

A Brief Introduction to Machine Learning

The developments that have driven the recent wave of excitement around artificial intelligence have occurred in a specific subfield of research known as *machine learning*. Machine learning explores the creation of algorithms that improve through the processing of data. It is also the technology that has produced many of the most widely shared "deepfakes" which have received coverage in traditional media.

While the mathematics can be quite complex, the general intuition behind machine learning is easy to understand. Consider the process of programming a computer to recognize an animal like a cat in an image. One approach is to explicitly input a set of rules for differentiating cats from other objects that might appear. For instance, one might require that the computer look for particular fur patterns and feline facial features. This method is known in the field as *feature engineering*. In feature engineering, expert engineers articulate a defined set of rules that a machine can follow to achieve an expected result.

Machine learning takes a different approach. Machine learning allows the computer to "teach" itself how to accomplish a given task, rather than simply giving the machine a set of instructions to follow. First, engineers and researchers compile a large dataset known as a *training corpus*. For a cat recognition algorithm, this would entail aggregating a large number of images of cats that are tagged as containing cats. This data is then processed by a learning algorithm. The result of a successful training process is a piece of software referred to as a trained *model*. The cost and resource requirements vary wildly depending on the specific machine learning application. One standard processor used in industry is Nvidia's Tesla V100 GPU processor, which retails for several thousand dollars US. One recent

paper producing state-of-the-art deepfakes used eight such processors. If properly configured, the model "learns" to associate certain visual patterns in images with the tag "cat". This allows it to correctly identify objects in novel images that were not part of its original training corpus.

Given sufficient data and the right algorithms, machine learning models are able to significantly outperform systems constructed on the feature engineering paradigm. This result is thought to be due, in part, to computers being able to pick up on subtle nuances that humans may not see or may be unable to articulate in a feature engineering context. In certain domains, like computer vision, machine learning approaches have steadily outperformed and replaced older feature engineering techniques.[4]

A final nuance is that machine learning is not one thing. There are many different types of models that can be leveraged to perform machine learning. A great deal of work in the technical research field focuses on identifying which models are best at solving specific kinds of problems. "Deep neural networks"—one kind of model—have proven useful in creating high-performance systems in many different domains. Machine learning using these neural networks, sometimes referred to as "deep" learning, has allowed computers to achieve significant improvements in tasks, ranging from language translation to playing the game of Go.[5]

Producing the Deepfake

Deepfakes emerge from this basic approach to teaching machines. A machine learning model contains some level of understanding of the task that it is attempting to achieve. In order to successfully recognize a cat in an image, a machine learning model needs to have some understanding of what a cat looks like. Machine learning researchers refer to this understanding as a "representation." These representations can be quite limited; an image recognition model trained on images of cats may only be able to associate the visual appearance of a cat with the label "cat." It will not necessarily know that a cat is an animal, the habits of a cat, or that a cat is biologically related to other animals like lions or tigers.

Representations can be used to generate synthetic data which resembles the data that the model was originally trained upon. Models trained on images of cats can be designed to subsequently output new examples of "cats" which approximate those images. This is a useful technique for researchers since it allows them to gain a better understanding of precisely what a model has learned during the training process.

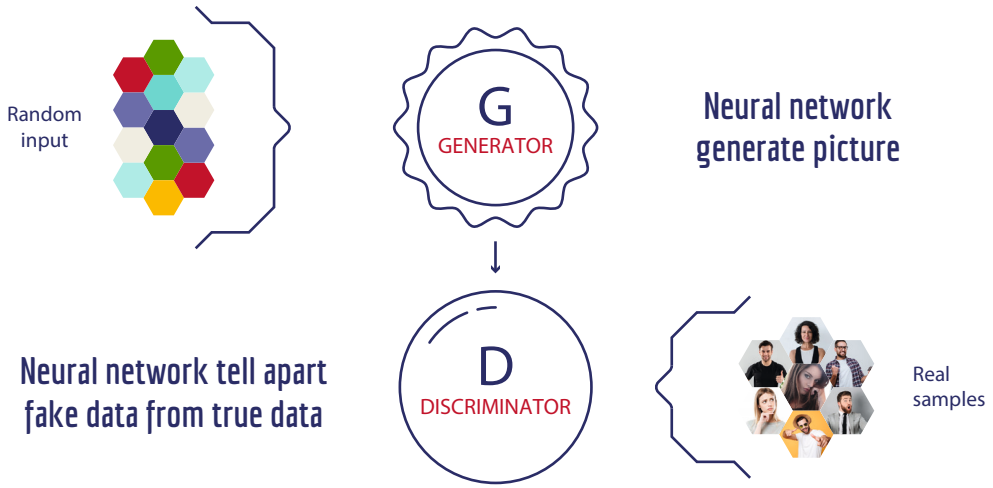
This method is at the core of how "deepfakes" are produced. Deep neural networks can acquire particularly rich representations of their training data, allowing them in turn to produce high-quality imitations that look strikingly like the real article.

GENERATIVE MODELS AND SAMPLING

The quality of these fakes has also been improved through a clever technique known as generative adversarial networks, or GANs. This method pairs two machine learning algorithms together: a generative network and a discriminative network. The discriminative network attempts to distinguish between real versions of its training data and fake versions of its training data. To take the earlier example, this might be a model which attempts to differentiate photos of real cats from synthetic, simulated photos of cats. The generative network in turn attempts to "fool" the discriminator by producing fakes that it accepts as genuine.

During the training process, these two machine learning algorithms compete with one another, with the generator working to produce faked inputs that a discriminator is unable to distinguish from the real article. The outcome of the competition is used to train each network, such that the discriminator becomes better at detecting fakes, and the generator becomes better at creating fakes. If successful, the generator can eventually produce synthetic media that is challenging for human observers to identify as fake.

GANs have produced some of the deepfakes which have been most widely shared beyond the research community in traditional media and across social media. One notably high-quality example is a manipulation of a video clip in which American actor and impressionist Bill Hader recounts to a talk-show host a conversation he had with two other actors, Tom Cruise and Seth Rogen. The manipulation shifts Hader's features subtly as he impersonates Cruise and Rogen, their features phasing in and out of



GAN Process design

Hader's almost seamlessly.[6] It also includes a 2018 demonstration from chipmaker NVIDIA showing that GANs can be used to create vivid, high-resolution synthetic faces.[7] GANs are also the technology behind "Everybody Dance Now", a 2019 research paper demonstrating a model for producing "do as I do" synthetic videos.[8] In the demonstration, researchers are able to use a video of a professional dance routine to produce a synthetic video of a graduate student performing the same motions.[9]

This technical background lays the groundwork for assessing the current state of play in deepfake technology. Part II discusses how these research advances are translating into real world uses.

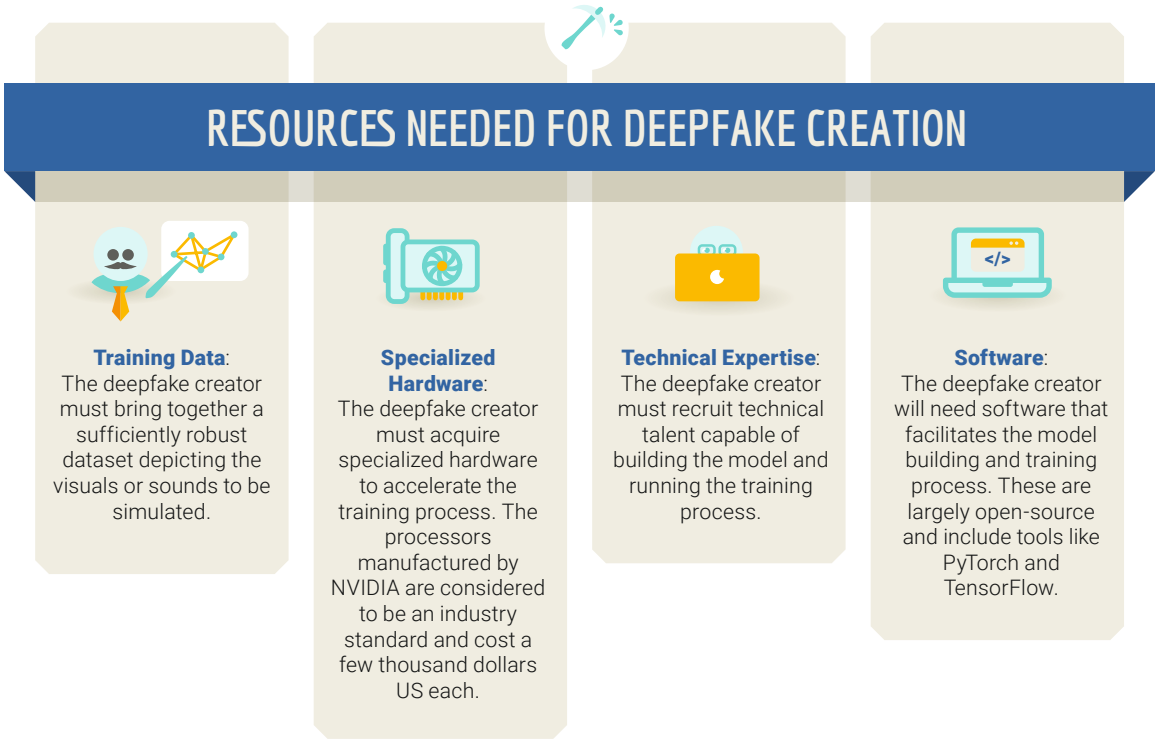
THE EXISTING STATE OF PLAY

Researchers have been able to show that machine learning can be used to produce extremely convincing fakes. But, as in other domains, a lab demonstration can be a far cry from how a technology is ultimately used in the real world. The movement of deepfakes from the lab into the mainstream is shaped by three major trends: *democratization, improving detection,* and *the limits of artificial intelligence.*

Trend 1: Democratization

Training a machine learning model can be a resource-intensive process. Domain experts must design the model, a sufficiently robust training corpus must be assembled, and specialized hardware needs to be acquired. Graphics Processing Units, or GPUs, are the standard processor used for machine learning applications and can cost a few thousand dollars each. Researchers have used up to eight of these processors in state-of-the-art deepfake generation.[10]

This may be in part why deepfakes have not become as ubiquitous as some have



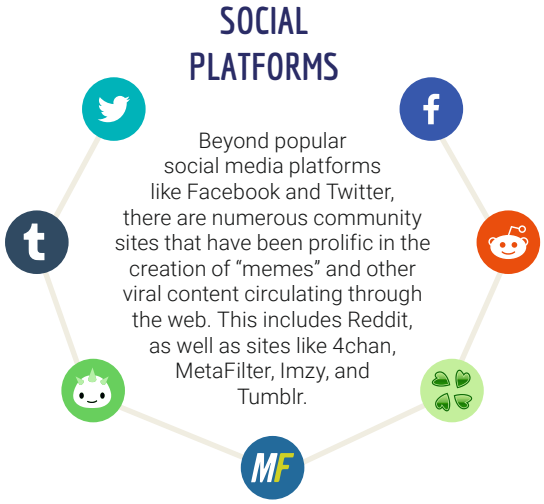
expected in the political domain—the benefit that a disinformation campaign stands to reap from using the technology does not yet sufficiently outweigh the costs.

This may be particularly true when "cheapfakes"—manipulated media created through decidedly low-tech means—have proven to be highly effective in spreading false narratives. This includes videos that have been sped up or slowed down to give an erroneous impression that an individual is drunk or has a medical problem. U.S. Speaker of the House Nancy Pelosi is a frequent target of this style of disinformation campaign,[11] or simply images taken from elsewhere and given a misleading caption[12] – as when a photograph of a Somali woman, taken in 1978, was circulated with a caption identifying the woman in the image as U.S. Congressmember Ilhan Omar (who was born in 1982).[13]

Disinformation actors may prefer these tried-and-true methods when compared with the resource requirements and special expertise needed to create a comparable deepfake. As the concept of "deepfakes" grows familiar to general audiences, viewer skepticism may also work to reduce the effectiveness of disinformation campaigns that rely on this technology alone.

This balance of costs and benefits will likely change going forward. While training a high-performance machine learning system can be a complex technical task, the resulting model is simply a small piece of software. The knowledge to build these systems is well within the reach of even moderately well-resourced actors. Governments, companies, and intelligence services all currently use machine learning. Trained models are easy for non-specialist software engineers to use and low cost to distribute across the web. Models can also be integrated into user-friendly apps, making it easy for laypeople to quickly produce deepfakes without any technical knowledge.

It is reasonable to expect that the technology to accomplish relatively simple deepfake media manipulations will become increasingly accessible. This has already begun to happen. The "face swap"—a standard video transformation in which an existing face is replaced by a new face selected by the manipulator—provides perhaps the clearest example of this trend. In 2017, a community was launched on the social platform Reddit to share synthetic pornography generated through face swap machine learning models.[14] These models were eventually integrated into a program called FakeApp, which made it easy for anyone to create face swap videos of their own. The creator wanted to make deepfakes "available to people without a technical background or programming experience," envisioning that the app would eventually allow "prospective users [to] simply select a video on their computer, download a neural network correlated to a certain face…and swap the video with a different face with the press of one button."[15]



After considerable controversy about the use of these apps to produce nonconsensual explicit imagery, FakeApp was taken down and the deepfakes community was banned on Reddit.[16] However, these actions have been unsuccessful in blocking public access to the technology. FakeApp has since been superseded by an open-source version of the app called Faceswap, and the code for accomplishing face swaps is now openly available as a software package called DeepFaceLab.[17]

Deepfake technologies have also been integrated into consumer applications and artistic projects. Zao, launched by the Chinese company Momo in 2019, enables users to create and share "face swap" videos with one another. Thispersondoesnotexist.com is an art project that allows users to generate realistic looking images of fake people.

FakeApp, Zao, and Thispersondoesnotexist.com represent a movement from specialized technology to commodified apps that other deepfake technologies will follow. It only requires a relatively small number of technically sophisticated actors to make a deepfake creation technique widely available to non-technical users. And, as with other forms of software, it can be difficult or impossible to control the spread of these apps once released to the public.

Trend 2: The Improving State of Detection

The rapid transmission of deepfake technologies from the research lab to user-friendly app might be a cause for despair. This trend might point towards a future scenario in which society is awash in deepfakes, unable to effectively differentiate truth from falsehood. However, the democratizing trend in the technology must be considered alongside an equally important trend: the improving state of deepfake detection. Deepfakes are a technology that can be used for many different purposes. Political criticism/satire is an activity, which might incorporate deepfakes to achieve a certain purpose.

The rising concern around deepfakes and their potential use by malicious actors has spurred governments and private industry into action. Major social media channels including Facebook, Twitter, and Reddit have rolled out new policies forbidding the use of deepfakes for disinformation purposes.[18] Companies[19] have also invested resources in advancing the state of the art in detection

methodologies. In 2019, Facebook, Amazon, Microsoft, and the civil society organization Partnership on AI, launched a public challenge in deepfake detection, seeking to encourage people to "build innovative new technologies that can help detect deepfakes and manipulated media."[20] Google has released datasets of synthetic speech and video to aid researchers in developing new detection methods and creating standard benchmarks in the community.[21] Google Jigsaw has also released Assembler, an experimental platform that automates a range of media forensics tasks.[22]

Governments have also taken action to accelerate research in the space. At the time of writing, this includes the Media Forensics program at the Defense Advanced Research Projects Agency (DARPA), which "brings together world-class researchers to attempt to level the digital imagery playing field, which currently favors the manipulator, by developing technologies for the automated assessment of the integrity of [media]… and integrating these in an end-to-end media forensics platform."[23] The Identifying Outputs of Generative Adversarial Networks (IOGAN) Act—currently pending in the U.S. Congress—would direct government research resources towards advancing the detection of manipulated media.[24]

Research efforts focused on detecting deepfakes are showing fruit. Detection algorithms are able to identify subtle warping and other artifacts produced by GANs and other generative models.[25] Other approaches, drawing on existing techniques in the field of media forensics, attempt to detect physiological signals like blinking and skin flushing that faces in deepfake videos frequently lack[26]. As in the creation of deepfakes, these research developments appear likely to be integrated into user-friendly software that will aid in deepfake detection. Deeptrace, an Amsterdam-based startup using "deep learning and computer vision for detecting and monitoring AI-generated synthetic videos," is likely the first of many companies that will offer products and services in this space.[27]

This does not mean that deepfakes are a solved problem. There remain a number of challenges in deepfake detection that the research community has not yet been able to surmount. Presently, one of the biggest problems is that machine learning algorithms can generalize poorly. This means that they learn to be successful at accomplishing a given task, but only for data that closely resembles the original training corpus. Consider the example of the machine learning model for detecting cats in images. If the model is trained purely on photos of cat faces, it may be unable to detect the animal in photographs in which the face of the cat does not appear or is otherwise obscured. This makes the model less useful for detecting cats.

This is a significant problem in the context of deepfake detection. Many of the most successful detection methodologies themselves rely on machine learning. These

algorithms train on examples of genuine media and synthetic media to differentiate between them. However, these models have been found to fail when encountering synthetic media generated through novel means that were not incorporated into the original training corpus.[28] This means that detection algorithms will be of limited usefulness in situations where malicious actors engineer entirely new synthetic media models rather than using open source, off-the-shelf ones. Researchers are working to surmount this obstacle, but it remains a persistent issue on the disinformation defense side of the equation.

Trend 3: Limits of Artificial Intelligence

Despite great advances in the field of machine learning, the technology remains limited in significant ways. These will impact the extent to which deepfakes can be leveraged by malicious actors to influence public discourse and spread disinformation.

For one, machine learning systems remain extremely limited in their understanding of context. As discussed above, a generative model for creating images of cats may have an internal representation of what a cat looks like. However, this representation may not incorporate a broader understanding of the habits of a cat, or even of how cats move.

This has practical implications for the production of deepfakes. For example, a good deepfake generator may have a robust visual representation of what a political leader might look like when speaking, and even what this leader sounds like. But ultimately, these models are flat "puppets" devoid of a broader understanding of context. For the foreseeable future, many of these models will rely on human writers to develop a script which will be believable and persuasive to the audience that eventually sees it. The media manipulator will also need to make curatorial choices about what to depict in a deepfake: Where will a political leader be shown speaking? Will they appear with anyone else? What motions will they make?

While deepfakes may give malicious actors the ability to create synthetic media that fools the eye, their ultimate persuasive strength will continue to rely on the ingenuity—or incompetence—of the human that assembles the hoax. Deepfakes therefore remain vulnerable to many of the same investigative techniques that have uncovered fakes in past decades. Investigators might seek corroborating evidence that supports or discredits what is depicted in a suspect video, or seek to identify how a video originated and was distributed. Machine learning models will remain relatively weak in simulating these contextual details, giving forensic experts a fighting chance.

Secondly, machine learning remains extremely dependent on data. As a general matter, the lack of a relevant training corpus prevents a learning algorithm from

generating a high-performance model. One of the reasons that celebrities and political leaders have been such popular subjects for deepfake generation by researchers is that there exists plentiful training media depicting these figures.

This imposes some practical limits on a malicious actor seeking to use generative models. Machine learning models cannot be used to simulate anyone and anything. Instead, GANs and other models will be most successful where there exists a large corpus of available data for the algorithm to train on. The quality of a given deepfake will accordingly decline as less and less data is available. The practical cost of acquiring data will accordingly limit the kinds of events and individuals that can be persuasively depicted in a deepfake.

ASSESSING THE FUTURE OF DEEPFAKES

Various pressures will influence how deepfakes are used in practice by disinformation perpetrators. The ongoing democratization of the technology will expand access going forward, enabling more and more actors to be able to create deepfakes. At the same time, the improving state of detection and the persistent limits of machine learning will also temper the impact this technology has on public discourse.

How will these various trends intersect? What are the prospects for deepfakes in the near-term and long-term? Part III addresses these trends, assesses the overall risk, and proposes some recommendations for addressing the threat posed by deepfakes.

The Future Threat Landscape

The commodification of generative models into apps that are easy for non-technical users to adopt may lead to the assumption that society will be saturated with deepfakes. While we should expect to see the number of deepfakes circulating through the web to increase over time, the effects may not be as dire as might initially seem to be the case. This is due to the trends that currently exist in research on deepfake detection.

Deepfake detection systems are most successful when researchers have access to many examples of synthetic media produced by a specific model. These examples are used as a training corpus, which in turn can be used to create a detection system capable of detecting the signatures of that generative model.[29]

This introduces an interesting dynamic to the landscape of deepfake creation and detection. Machine learning models that are widely shared and integrated into apps are also likely to produce significant bodies of synthetic media circulating through the web. This simultaneously makes these commodified techniques precisely those which are most susceptible to automated detection. These detection algorithms can be used to enhance the enforcement of policies on online platforms, create standalone apps that help to inform users online, and be leveraged by civil society in countering disinformation efforts. The creation of simple deepfakes like the "face

swap" will become easier, but detecting and rooting them out will become easier as well. This will limit their impact over time.

This is not to argue that deepfakes do not pose a threat. However, the threat from deepfakes is somewhat narrower than is sometimes implied.[30] Due to the overfitting problem, the deepfakes most able to evade automated detection are those created by novel generative models. In these cases, debunking deepfake images, audio, and video may require investigation by domain experts who will need to use contextual clues and other strategies to detect the manipulation. This slower, manual process expands the opportunity that this media has to spread throughout the web and can make it difficult to render a definitive forensic analysis.

Sophisticated, well-resourced disinformation campaigns, like those launched during the 2016 U.S. presidential race, may therefore represent the most significant deepfake threat. These campaigns have the resources to construct tailored machine learning models for their own purposes. They also can invest in collecting unique training data to depict events and individuals otherwise challenging to generate with publicly available images and video. These deepfakes might be held in reserve until a crucial moment, enabling a malicious actor to avoid publishing examples that might be used to improve a detection system.

This scenario starkly contrasts with one in which the primary threat is the democratization of deepfake technologies. Instead, the real threat might lie in a small number of highly-sophisticated fakes created and launched by well-organized disinformation campaigns. The technology is well within the reach of states, corporations, and even moderately well-resourced private individuals. These will be the most challenging to prepare for in advance, to detect once they have been launched, and to refute once identified.

The Deepfake Distraction?

Given that the threat is narrower than commonly portrayed, it is possible that deepfakes may serve as a distraction from the deeper problems of disinformation and misinformation on two counts.

First, deepfakes may distract from other deployments of machine learning by disinformation campaigns that may be as, or even more impactful, than deepfakes. Machine learning has a wide range of potential applications in this domain. Malicious actors might leverage advances in conversational agents to produce swarms of false identities able to interact believably with real users. Disinformation efforts might use machine learning to model social behavior, allowing them to target their persuasive efforts far more effectively. Focusing on deepfakes as the sole or primary means for which machine learning will be used may miss more significant but less vivid threats posed by the technology.

Secondly, deepfakes may distract from the psychological and sociological aspects of disinformation and misinformation. As the widespread sharing of crude edited photographs and other "cheapfakes" shows, the level of realism or the sharpness of an image does not determine the successful spread of a false narrative. The voluminous literature on these topics reveals the many motivations—from a desire to show group identity, to a pre-existing belief in the narrative reinforced by a piece of media—that determine whether a faked image or video is believed, and whether an individual decides to share it more broadly.[31] It is unclear whether the striking realism offered by a deepfake will alone make it a significant tool for manipulating the public.

Moreover, addressing deeper "demand-side" factors in the spread of disinformation and misinformation may lead to more robust defenses against these threats. The aim should be to make audiences generally less receptive to false narratives over time, rather than attempt to combat a specific means of generating false narratives. Even the most effective deepfake detection system will be vulnerable to disinformation perpetrators simply adopting a different means of distributing their false narrative.

Deepfakes represent a novel addition to the disinformation perpetrator's toolkit, but it is unclear that the technology represents a true game-changer. Instead, the risks presented by the technology will be narrower, and the responses should be targeted to the threat.

RECOMMENDATIONS

Based on this analysis, there are four proposals that would make a significant difference in combatting the threat posed by deepfakes.

Bridge Media Forensics and Strategic Communications

The research field of media forensics has developed a wide range of tools for identifying manipulated media of many kinds. This includes recent work on detecting deepfakes, as well as research examining digital artifacts left by other forms of image and video modification. One example includes the close examination of digital traces that are left by photo and video editing software.[32] These are important techniques that should be brought into the day-to-day work of strategic communications. However, by and large these communities are not closely intertwined with one another.

The strategic communications profession should build collaborations and connections with the media forensics community. This will enable communications experts to bring the latest tools to bear in their day-to-day work. It would also create a productive channel of communication to ensure that the research community has access to the latest knowledge about the emerging media manipulation techniques and hard problems being observed "in the field."

Accelerate Detection Democratization

Deepfake detection should be commoditized in order to keep up with the ongoing commodification of deepfake creation. While a few startups, like the Amsterdam-based DeepTrace Labs, are beginning to offer products and services, the transmission of research findings to practical tools should be accelerated as much as possible. Deepfakes will be less able to spread in an online environment in which citizens, civil society organizations, online platforms, and governments all have easy access to state-of-the-art detection tools. Funding startups working on these problems and hosting public challenges to encourage work in the area will both help to democratize detection.

Detection through automated software systems is just one part of this. There are many other techniques for interrogating and investigating the veracity of suspect media that do not rely on detection algorithms.[33] Scaling media literacy programs and training in identifying fake media

will help to build public resilience against deepfakes and help to slow their influence when they are used by malicious actors.

Invest in Research on the Cognitive Dimension of Deepfakes

There remain a number of important unknowns about how audiences understand and respond to high-quality synthetic media. Does the level of video and image quality determine how believable the media is? How much of a factor is it? Does existing knowledge of deepfakes make an individual less willing to believe media in general? If so, how significant is this effect?

The machine learning field remains focused on the narrower technical problems of detecting the use of deepfakes. It will be important to accompany this work with research into the psychological and sociological aspects of deepfakes. This would include investments in public polling to understand how widespread knowledge of deepfakes is and to what extent different demographics are on guard to the risk posed by these technologies. This effort would also include supporting research examining how and why different kinds of deepfakes are believed and shared by their audiences. These finding would help to ground the public policy debate around these technologies and advance informed responses to the threat that they pose.

Invest in Next Generation Detection Techniques

The creation and detection of deepfakes will likely evolve towards a cat-and-mouse game. Detection techniques for identifying deepfakes will improve, which will encourage malicious actors to improve their methods of deepfake creation. This will set off another cycle of investment and research effort in detection.

To break this cycle, investments should be made in advancing next generation technologies that might give anti-disinformation efforts a "leap-frogging" advantage over disinformation campaigns seeking to use deepfakes in their operations. Currently, researchers are exploring tamper-proof watermarks that would be attached to images captured authentically by digital cameras.[34] This might allow faster verification of whether an image or a video was generated synthetically by a machine learning model. There has also been research into making datasets "radioactive," such that models trained on them will produce fakes which reveal that they have used this data.[35] This might make it harder for disinformation campaigns to rely on public data in training their deepfake models. These opportunities might change the strategic balance surrounding deepfakes, allowing detection efforts to gain a significant lead over deepfake creators.

CONCLUSION

Disinformation campaigns pose a significant threat to democratic processes, public health, and market stability. As numerous case studies over the past few years attest, these malicious efforts can be nimble, well-resourced, and technologically savvy. It seems likely that these campaigns will at some point harness the latest developments in machine learning for destructive ends. Indeed, what is more surprising is that the technology has not seen greater use to date. This is poised to change as tools for creating deepfakes commodify and become increasingly accessible over time.

However, the dramatic demonstrations of deepfake generation published by researchers should not lead to the conclusion that the technology is some ultimate weapon of mass disinformation. Deepfake detection is evolving alongside deepfake creation. It is more likely that detection and creation will be headed towards a long-term stalemate: as one improves, so does the other.

However, detection will be most effective when many samples of a given deepfake creation method are available. These samples can be used by researchers to train and improve detection algorithms. This suggests that the threat from commodified deepfakes—simple transformations of media like the face swap—will pose less of a threat over time as their usage becomes more widespread. There will simply be more samples available to train effective detection systems.

The bigger threat for identifying deepfakes will come from tailored models built by sophisticated actors and released at critical points to produce the greatest harm. These are the points where detection systems are likely to fail, and where synthetic media can be crafted to have the greatest resonance with the public.

Anti-disinformation initiatives should tailor their efforts to deal with these threats. High impact opportunities include building connections with the technical media forensics community, investing in the democratization of detection technologies, supporting research on the psychological dimensions of deepfakes, and investing in next generation detection techniques. Putting these efforts into motion will play a major role in ensuring that malicious uses of deepfakes encounter an inhospitable environment online where it is possible to debunk and refute these fakes as quickly as they emerge.

Even in the midst of these efforts, it is important that deepfakes do not become a distraction. Faked images and video are

simply one tool among many in the hands of malicious actors. Overinvestment in countering this cutting-edge tool may simply encourage media manipulation campaigns to adopt alternative tools that are equally effective in eroding trust in the overall information environment. Ultimately, resilience against online disinformation will depend not only on the ability to harness technology, but the ability to harness social and psychological forces, as well.

Endnotes

1 See, e.g., NOVA, "Deepfake Videos Are Getting Terrifyingly Real", YouTube.

2 Regina Rini, "Deepfakes Are Coming. We Can No Longer Believe What We See", *New York Times*, June 10, 2019,

3 Robert Chesney and Danielle Keats Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security", University of Texas Law, Public Law Research Paper No. 692, July 21, 2018.

4 Dave Gershgorn, "The data that transformed AI research—and possibly the world", *Quartz*, July 26, 2017.

5 See Gideon Lewis-Kraus, "The Great A.I. Awakening," *New York Times*, December 14, 2016, (machine learning for translation); DeepMind, AlphaGo, (machine learning for Go).

6 Ctrl Alt Face, "Bill Hader channels Tom Cruise [DeepFake]," YouTube.

7 Tero Karras et al., "Progressive Growing of GANs for Improved Quality, Stability, and Variation." Preprint, submitted October 27, 2017.

8 Caroline Chan et al., "Everybody Dance Now." Preprint, submitted August 22, 2018.

9 Ibid.

10 See Karras et al., "Progressive Growing" (leveraging eight NVIDIA Tesla V100 GPUs, a common processor for machine learning applications).

11 See, e.g. "Pelosi videos manipulated to make her appear drunk are being shared on social media", *Washington Post,* May 23, 2019.

12 Snopes, "Out-of-Context Photos Are a Powerful Low-Tech Form of Misinformation", *Snopes*, February 15, 2020.

13 Dan Evon, "Did This Picture Show U.S. Rep. Ilhan Omar in 'Jiahd Academy'", *Snopes,* August 21, 2019,

14 Samatha Cole, "AI-Assisted Fake Porn Is Here and We're All Fucked," *Motherboard*, December 11, 2017.

15 Ibid.

16 Megan Karokhmanesh, "Deepfakes are Disappearing from Parts of the Web, But They're Not Going Away", *The Verge*, February 9, 2018.

17 "DeepFaceLab", Github.

18 Monika Bickert, "Enforcing Against Manipulated Media," Facebook, January 6, 2020,; Del Harvey, "Help us shape our approach to synthetic and manipulated media", Twitter, November 11, 2019,; Jay Peters, "Reddit bans impersonation on its platform", *The Verge*, January 9, 2020.

19 This video covers some of the ongoing detection efforts in this space: https://www.youtube.com/watch?v=4YpoYvhVmDw

20 Mike Schroepfer, "Creating a dataset and a challenge for deepfakes", *Facebook Artificial Intelligence* (blog), September 5, 2019.

21 Nick Dufour and Andrew Gully, "Contributing Data to Deepfake Detection Research", *Google AI* (blog), September 24, 2019.

22 Google Jigsaw, Assembler.

23 DARPA, "Media Forensics (MediFor)".

24 Brandi Vincent, "Bill to Combat Deepfakes Passes House Committee", *NextGov*, September 26, 2019.

25 See Luisa Verdoliva, "Media Forensics and DeepFakes: an overview." Preprint, submitted January 18, 2020.

26 Ibid.

27 "Deeptrace", Deeptrace Labs.

28 Rayhane Mama and Sam She, "Towards Deepfake Detection That Actually Works", *Dessa* (blog), November 24, 2019.

29 Andreas Rossler, "FaceForensics++: Learning to Detect Manipulated Facial Images."Preprint, submitted January 25, 2019.

30 See, e.g, Alec Radford et al., "Better Language Models and Their Implications", *OpenAI*, February 14, 2019, (limiting publication of a generative model due to fears that distribution would lead to misuse).

31 See Joshua Tucker et al., "Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature", Hewlett Foundation (March 2018),.

32 Luisa Verdoliva, "Media Forensics and DeepFakes: an overview." Preprint, submitted January 18, 2020, 5-6,.

33 See, e.g., First Draft, "First Draft launches its online verification training course," *First Draft*, October 11, 2017, https://firstdraftnews.org/latest/course-launches/.

34 Lily Hay Newman, "To Fight Deepfakes, Researchers Built a Smarter Camera," *Wired*, May 28, 2019, https://www.wired.com/story/detect-deepfakes-camera-watermark/.

35 Alexandre Sablayrolles, Matthijs Douze, and Hervé Jégou, "Using 'radioactive data' to detect if a dataset was used for training," *Facebook AI Blog*, February 5, 2020, https://ai.facebook.com/blog/using-radioactive-data-to-detect-if-a-data-set-was-used-for-training/.

Prepared and published by the
NATO STRATEGIC COMMUNICATIONS
CENTRE OF EXCELLENCE

The NATO Strategic Communications Centre of Excellence (NATO StratCom COE) is a NATO accredited multi-national organisation that conducts research, publishes studies, and provides strategic communications training for government and military personnel. Our mission is to make a positive contribution to Alliance's understanding of strategic communications and to facilitate accurate, appropriate, and timely communication among its members as objectives and roles emerge and evolve in the rapidly changing information environment.

Operating since 2014, we have carried out significant research enhancing NATO nations' situational awareness of the information environment and have contributed to exercises and trainings with subject matter expertise.

www.stratcomcoe.org | @stratcomcoe | info@stratcomcoe.org