Deepfakes and the New Disinformation War

The Coming Age of Post-Truth Geopolitics

By Robert Chesney and Danielle Citron

A picture may be worth a thousand words, but there is nothing that persuades quite like an audio or video recording of an event. At a time when partisans can barely agree on facts, such persuasiveness might seem as if it could bring a welcome clarity. Audio and video recordings allow people to become firsthand witnesses of an event, sparing them the need to decide whether to trust someone else's account of it. And thanks to smartphones, which make it easy to capture audio and video content, and social media platforms, which allow that content to be shared and consumed, people today can rely on their own eyes and ears to an unprecedented degree.

Therein lies a great danger. Imagine a video depicting the Israeli prime minister in private conversation with a colleague, seemingly revealing a plan to carry out a series of political assassinations in Tehran. Or an audio clip of Iranian officials planning a covert operation to kill Sunni leaders in a particular province of Iraq. Or a video showing an American general in Afghanistan burning a Koran. In a world already primed for violence, such recordings would have a powerful potential for incitement. Now imagine that these recordings could be faked using tools available to almost anyone with a laptop and access to the Internet—and that the resulting fakes are so convincing that they are impossible to distinguish from the real thing.

Advances in digital technology could soon make this nightmare a reality. Thanks to the rise of "deepfakes"—highly realistic and difficult-to-detect digital manipulations of audio or video—it is becoming easier than ever to portray someone saying or doing something he or she never said or did. Worse, the means to create deepfakes are likely to proliferate quickly, producing an ever-widening circle of actors capable of deploying them for political purposes. Disinformation is an ancient art, of course, and one with a renewed relevance today. But as deepfake technology develops and spreads, the current disinformation wars may soon look like the propaganda equivalent of the era of swords and shields.

DAWN OF THE DEEPFAKES

Deepfakes are the product of recent advances in a form of artificial intelligence known as "deep learning," in which sets of algorithms called "neural networks" learn to infer rules and replicate patterns by sifting through large data sets. (Google, for instance, has used this technique to develop powerful image-classification algorithms for its search engine.) Deepfakes emerge from a specific type of deep learning in which pairs of algorithms are pitted against each other in "generative adversarial networks," or GANS. In a GAN, one algorithm, the "generator," creates content modeled on source data (for instance, making artificial images of cats from a database of real cat pictures), while a second algorithm, the "discriminator," tries to spot the artificial content (pick out the fake cat images). Since each algorithm is constantly training against the other, such pairings can lead to rapid improvement, allowing GANS to produce highly realistic yet fake audio and video content.

This technology has the potential to proliferate widely. Commercial and even free deepfake services have already appeared in the open market, and versions with alarmingly few safeguards are likely to emerge on the black market. The spread of these services will lower the barriers to entry, meaning that soon, the only practical constraint on one's ability to produce a deepfake will be access to training materials—that is, audio and video of the person to be modeled—to feed the GAN. The capacity to create professional-grade forgeries will come within reach of nearly anyone with sufficient interest and the knowledge of where to go for help.

Deepfakes have a number of worthy applications. Modified audio or video of a historical figure, for example, could be created for the purpose of educating children. One company even claims that it can use the technology to restore speech to individuals who have lost their voice to disease. But deepfakes can and will be used for darker purposes, as well. Users have already employed deepfake technology to insert people's faces into pornography without their consent or knowledge, and the growing ease of making fake audio and video content will create ample opportunities for blackmail, intimidation, and sabotage. The most frightening applications of deepfake technology, however, may well be in the realms of politics and international affairs. There, deepfakes may be used to create unusually effective lies capable of inciting violence, discrediting leaders and institutions, or even tipping elections.

Social media will be fertile ground for circulating deepfakes, with explosive implications for politics.

Deepfakes have the potential to be especially destructive because they are arriving at a time when it already is becoming harder to separate fact from fiction. For much of the twentieth century, magazines, newspapers, and television broadcasters managed the flow of information to the public. Journalists established rigorous professional standards to control the quality of news, and the relatively small number of mass media outlets meant that only a limited number of individuals and organizations could distribute information widely. Over the last decade, however, more and more people have begun to get their information from social media platforms, such as Facebook and Twitter, which depend on a vast array of users to generate relatively unfiltered content. Users tend to curate their experiences so that they mostly encounter perspectives they already agree with (a tendency heightened by the platforms' algorithms), turning their social media feeds into echo chambers. These platforms are also susceptible to so-called information cascades, whereby people pass along information shared by others without bothering to check if it is true, making it appear more credible in the process. The end result is that falsehoods can spread faster than ever before.

These dynamics will make social media fertile ground for circulating deepfakes, with potentially explosive implications for politics. Russia's attempt to influence the 2016 U.S. presidential election—spreading divisive and politically inflammatory messages on Facebook and Twitter—already demonstrated how easily disinformation can be injected into the social media bloodstream. The deepfakes of tomorrow will be more vivid and realistic and thus more shareable than the fake news of 2016. And because people are especially prone to sharing negative and novel information, the more salacious the deepfakes, the better.

DEMOCRATIZING FRAUD

The use of fraud, forgery, and other forms of deception to influence politics is nothing new, of course. When the USS *Maine* exploded in Havana Harbor in 1898, American tabloids used

misleading accounts of the incident to incite the public toward war with Spain. The anti-Semitic tract *Protocols of the Elders of Zion,* which described a fictional Jewish conspiracy, circulated widely during the first half of the twentieth century. More recently, technologies such as Photoshop have made doctoring images as easy as forging text. What makes deepfakes unprecedented is their combination of quality, applicability to persuasive formats such as audio and video, and resistance to detection. And as deepfake technology spreads, an ever-increasing number of actors will be able to convincingly manipulate audio and video content in a way that once was restricted to Hollywood studios or the most well-funded intelligence agencies.

Deepfakes will be particularly useful to nonstate actors, such as insurgent groups and terrorist organizations, which have historically lacked the resources to make and disseminate fraudulent yet credible audio or video content. These groups will be able to depict their adversaries—including government officials—spouting inflammatory words or engaging in provocative actions, with the specific content carefully chosen to maximize the galvanizing impact on their target audiences. An affiliate of the Islamic State (or ISIS), for instance, could create a video depicting a U.S. soldier shooting civilians or discussing a plan to bomb a mosque, thereby aiding the terrorist group's recruitment. Such videos will be especially difficult to debunk in cases where the target audience already distrusts the person shown in the deepfake. States can and no doubt will make parallel use of deepfakes to undermine their nonstate opponents.

Deepfakes will also exacerbate the disinformation wars that increasingly disrupt domestic politics in the United States and elsewhere. In 2016, Russia's state-sponsored disinformation operations were remarkably successful in deepening existing social cleavages in the United States. To cite just one example, fake Russian accounts on social media claiming to be affiliated with the Black Lives Matter movement shared inflammatory content purposely designed to stoke racial tensions. Next time, instead of tweets and Facebook posts, such disinformation could come in the form of a fake video of a white police officer shouting racial slurs or a Black Lives Matter activist calling for violence.

Perhaps the most acute threat associated with deepfakes is the possibility that a well-timed forgery could tip an election. In May 2017, Moscow attempted something along these lines. On the eve of the French election, Russian hackers tried to undermine the presidential campaign of Emmanuel Macron by releasing a cache of stolen documents, many of them doctored. That effort failed for a number of reasons, including the relatively boring nature of the documents and the effects of a French media law that prohibits election coverage in the 44 hours immediately before a vote. But in most countries, most of the time, there is no media blackout, and the nature of deepfakes means that damaging content can be guaranteed to be salacious or worse. A convincing video in which Macron appeared to admit to corruption, released on social media only 24 hours before the election, could have spread like wildfire and proved impossible to debunk in time.

Deepfakes may also erode democracy in other, less direct ways. The problem is not just that deepfakes can be used to stoke social and ideological divisions. They can create a "liar's dividend": as people become more aware of the existence of deepfakes, public figures caught in genuine recordings of misbehavior will find it easier to cast doubt on the evidence against them. (If deepfakes were prevalent during the 2016 U.S. presidential election, imagine how much easier it would have been for Donald Trump to have disputed the authenticity of the infamous audiotape in which he brags about groping women.) More broadly, as the public becomes sensitized to the threat of deepfakes, it may become less inclined to trust news in general. And journalists, for their part, may become more wary about relying on, let alone

publishing, audio or video of fast-breaking events for fear that the evidence will turn out to have been faked.

DEEP FIX

There is no silver bullet for countering deepfakes. There are several legal and technological approaches—some already existing, others likely to emerge—that can help mitigate the threat. But none will overcome the problem altogether. Instead of full solutions, the rise of deepfakes calls for resilience.

Three technological approaches deserve special attention. The first relates to forensic technology, or the detection of forgeries through technical means. Just as researchers are putting a great deal of time and effort into creating credible fakes, so, too, are they developing methods of enhanced detection. In June 2018, computer scientists at Dartmouth and the University at Albany, SUNY, announced that they had created a program that detects deepfakes by looking for abnormal patterns of eyelid movement when the subject of a video blinks. In the deepfakes arms race, however, such advances serve only to inform the next wave of innovation. In the future, GANS will be fed training videos that include examples of normal blinking. And even if extremely capable detection algorithms emerge, the speed with which deepfakes can circulate on social media will make debunking them an uphill battle. By the time the forensic alarm bell rings, the damage may already be done.

A second technological remedy involves authenticating content before it ever spreads—an approach sometimes referred to as a "digital provenance" solution. Companies such as Truepic are developing ways to digitally watermark audio, photo, and video content at the moment of its creation, using metadata that can be logged immutably on a distributed ledger, or blockchain. In other words, one could effectively stamp content with a record of authenticity that could be used later as a reference to compare to suspected fakes.

In theory, digital provenance solutions are an ideal fix. In practice, they face two big obstacles. First, they would need to be ubiquitously deployed in the vast array of devices that capture content, including laptops and smartphones. Second, their use would need to be made a precondition for uploading content to the most popular digital platforms, such as Facebook, Twitter, and YouTube. Neither condition is likely to be met. Device makers, absent some legal or regulatory obligation, will not adopt digital authentication until they know it is affordable, in demand, and unlikely to interfere with the performance of their products. And few social media platforms will want to block people from uploading unauthenticated content, especially when the first one to do so will risk losing market share to less rigorous competitors.

A third, more speculative technological approach involves what has been called "authenticated alibi services," which might soon begin emerging from the private sector. Consider that deepfakes are especially dangerous to high-profile individuals, such as politicians and celebrities, with valuable but fragile reputations. To protect themselves against deepfakes, some of these individuals may choose to engage in enhanced forms of "lifelogging"—the practice of recording nearly every aspect of one's life—in order to prove where they were and what they were saying or doing at any given time. Companies might begin offering bundles of alibi services, including wearables to make lifelogging convenient, storage to cope with the vast amount of resulting data, and credible authentication of those data. These bundles could even include partnerships with major news and social media platforms, which would enable rapid confirmation or debunking of content.

Such logging would be deeply invasive, and many people would want nothing to do with it. But in addition to the high-profile individuals who choose to adopt lifelogging to protect themselves, some employers might begin insisting on it for certain categories of employees, much as police departments increasingly require officers to use body cameras. And even if only a relatively small number of people took up intensive lifelogging, they would produce vast repositories of data in which the rest of us would find ourselves inadvertently caught, creating a massive peer-to-peer surveillance network for constantly recording our activities.

LAYING DOWN THE LAW

If these technological fixes have limited upsides, what about legal remedies? Depending on the circumstances, making or sharing a deepfake could constitute defamation, fraud, or misappropriation of a person's likeness, among other civil and criminal violations. In theory, one could close any remaining gaps by criminalizing (or attaching civil liability to) specific acts—for instance, creating a deepfake of a real person with the intent to deceive a viewer or listener and with the expectation that this deception would cause some specific kind of harm. But it could be hard to make these claims or charges stick in practice. To begin with, it will likely prove very difficult to attribute the creation of a deepfake to a particular person or group. And even if perpetrators are identified, they may be beyond a court's reach, as in the case of foreign individuals or governments.

Another legal solution could involve incentivizing social media platforms to do more to identify and remove deepfakes or fraudulent content more generally. Under current U.S. law, the companies that own these platforms are largely immune from liability for the content they host, thanks to Section 230 of the Communications Decency Act of 1996. Congress could modify this immunity, perhaps by amending Section 230 to make companies liable for harmful and fraudulent information distributed through their platforms unless they have made reasonable efforts to detect and remove it. Other countries have used a similar approach for a different problem: in 2017, for instance, Germany passed a law imposing stiff fines on social media companies that failed to remove racist or threatening content within 24 hours of it being reported.

Yet this approach would bring challenges of its own. Most notably, it could lead to excessive censorship. Companies anxious to avoid legal liability would likely err on the side of policing content too aggressively, and users themselves might begin to self-censor in order to avoid the risk of having their content suppressed. It is far from obvious that the notional benefits of improved fraud protection would justify these costs to free expression. Such a system would also run the risk of insulating incumbent platforms, which have the resources to police content and pay for legal battles, against competition from smaller firms.

LIVING WITH LIES

But although deepfakes are dangerous, they will not necessarily be disastrous. Detection will improve, prosecutors and plaintiffs will occasionally win legal victories against the creators of harmful fakes, and the major social media platforms will gradually get better at flagging and removing fraudulent content. And digital provenance solutions could, if widely adopted, provide a more durable fix at some point in the future.

In the meantime, democratic societies will have to learn resilience. On the one hand, this will mean accepting that audio and video content cannot be taken at face value; on the other, it will mean fighting the descent into a post-truth world, in which citizens retreat to their private information bubbles and regard as fact only that which flatters their own beliefs. In short,

democracies will have to accept an uncomfortable truth: in order to survive the threat of deepfakes, they are going to have to learn how to live with lies.