DEEPFAKES

DETECTION AND PROTECTION IN THE ERA OF "FAKE NEWS"

PREPARED BY
Matt Essert T’20
CONSIDER A SCENARIO:

You are the CEO of a major packaged goods company. You've been doing well, and the stock price is reflecting it. You’re sitting down with a morning coffee to start your day and turn on CNBC to see yourself giving an interview. You don’t exactly remember giving this interview, but with the upcoming earnings announcement later this week, the past few days have been a bit of a whirlwind. You watch yourself, but you’re a bit confused. You’re not exactly sure why, but you look a bit … different.

“We’re sad to say that we have had to issue the largest recall in the history of our company,” you appear to say on the screen. “Every case of [your best-selling product] has been contaminated with poison and could kill anyone who consumes any of it. Do not buy our products! In fact, you probably shouldn’t buy anything we’ve made for at least the next 12 months.”

You’re numb. You never said this and none of this information is true. But you’re looking at yourself on national television telling the entire world that your products are poisonous and that they should not be purchased. Before you can think straight, your phone starts ringing non-stop, you get 100 texts, and 200 emails.

“What the hell is going on?”

For a moment, no one really knows. Before you can answer that question, your stock has plummeted. You pick up the phone to call your COO, CFO, CMO, and board – “I never said this!” you say. “But you’re saying it on TV!” they respond. “What is going on!”

Though it might seem like an episode of The Twilight Zone or Black Mirror, this situation, or something very much like it, may be a reality sooner than you might think.

AND THE ROOT CAUSE IS SIMPLE: DEEPFAKES.
Deepfakes, a portmanteau of “deep learning” and “fake news,” are synthetic media files in which machine learning and artificial intelligence technologies replace parts of existing video files with adulterated content to create a new video or photo files displaying what the deepfake’s creator intended. Simply put, deepfakes mash real and computer-generated video files to create something entirely new. And what is so scary about them is that they can look 100% real. While the technical details of how deepfakes are generated are not the focus of this article, suffice it to say advances in machine learning and neural network technologies have made the generation of deepfakes relatively simple.

As early as 1997, a project called the Video Rewrite program demonstrated how existing video footage could be modified to show someone mouthing the words from a different audio recording [1]. Thanks to several years of advances in computing power and machine learning research, 2017 brought the Synthesizing Obama project, in which researchers trained a recurrent neural network on hours of footage of former president Barack Obama speaking to produce a photorealistic “Faux-bama” that could appear to say anything its creators wanted [2]. Just a year later, American filmmaker and comedian/Obama impersonator Jordan Peele produced a PSA video in conjunction with BuzzFeed warning of the dangers of realistic-looking fake news [3]. And what was so disconcerting about that was that it was produced with a simple combination of Adobe After Effects and an AI face-swapping tool called FakeApp. That is to say, this wasn’t the result of complex, ultra-technical programming or hours of coding; it was something relatively simple that someone with a basic understanding of video editing, a few hours of time, and a good Obama impersonation could produce.

In fact, the evolution of this technology is moving at such a rapid pace that it will not be long before the steps to create such a video become even simpler. Hany Farid, a professor at the University of California, Berkeley with a joint appointment in Electrical Engineering & Computer Science and the School of Information, is an expert on digital forensics, image analysis, and human perception. He’s been studying the trajectory of deepfakes and expects them to become more commonplace as the technology becomes simpler and the use-cases more apparent. “We’re probably not far off from being to the point where, I wouldn’t say it’s point and click,” Professor Farid says, “But it’s starting to get there, where you can just say, ‘Okay, put this person in this video and have them say this.’ And I think that time is probably going to come. We just seem to be walking down that road right now” [4].

The Obama example was mostly a fun demonstration, but it did warn of what could happen when people are able to leverage the mistrust generated by the “fake news” epidemic. When distrust in the media and news jumps from not being able to believe what you read to not being able to believe what you see, we are in dire straits. To be sure, deepfakes are not about creating a perfect video that will always fool everyone. Often, it’s about being good enough for the objective at hand, especially given how quickly news can spread through wide-reaching social platforms. “I think the way you have to think about the risk is in the space where the deepfake only has to live for 24 or 48 hours,” Professor Farid says. “You don’t have to fool people all the time for a long period of time. It could be 24 or 48 hours before an election and a video of a candidate comes out saying some horrible things, or it’s a video of a President saying something racially insensitive and you have riots in the streets.”

In the history of photos and videos, image manipulation is not exactly new. However, much of the history of doctored images has been relatively easy to spot and debunk quickly. But with the rise of the Internet, social media, and near-instant information access and spreading, as the old saying goes, a lie can get halfway around the world before the truth gets its pants on. This means that people have realized that sometimes, when it comes to doctoring images or videos, “good enough” is often good enough. Rather than spend the time developing visually flawless and convincing deepfakes, people can simply produce what is known as a “cheap fake.” Cheap fakes require much less technical expertise and resources and can simply be the result of deceptive video editing such as speeding up or slowing down a video (i.e. to make it appear someone is slurring their words or punching someone with force when they are really just patting someone’s shoulder), or basic face swapping with rotoscoping technology (usually done to static images). In fact, it is helpful to think of cheap fakes and deepfakes not as two siloed options, but as two ends of a spectrum in which a variety of image adulterating can occur. This spectrum from a “Deepfakes and Cheap Fakes” report published in Data & Society is a helpful visualization of the complexity of different technologies that can be used in this arena and where their applications lie [27].

To be sure, while much simpler than deepfakes, they can be just as dangerous to the truth and to a business. While this article is mostly focusing on the threat of more advanced deepfakes, the same questions must be asked, principles must be applied, and preparations must be made to ready oneself for a world in which image veracity is increasingly called into question and the implications of a doctored image grow exponentially. In fact, rather than getting lost in semantics around what is and isn’t a deepfake, the focus should be on understanding the possibilities for and implications of any doctored images and videos and then thinking through how to prepare an organization to respond to issues caused by them.

[27] https://datasociety.net/library/deepfakes-and-cheap-fakes/
HOW ARE DEEPFAKES DEPLOYED?

One of the earliest instances of deepfake was in pornography, many of which surfaced online in 2017, most notably on Reddit and often imposing prominent celebrity’s faces into pornographic films [5]. Most social networks and image hosting websites – including Reddit, Twitter, Discord, Giphy, and PornHub – moved swiftly to remove any pornographic deepfake imagery from their sites [6, 7, 8, 9, 10].

On the political front, U.S. lawmakers have publicly recognized the potential threat of deepfakes when it comes to election and national security and overall political stability. With the rise of misinformation and “fake news” in the 2016 U.S. Presidential election, we seemed to have entered a world in which doctored imagery such as deepfakes could be weaponized to portray political rivals in certain lights, share false information, and sow overall discord in the political process. Shortly after a doctored video appearing to show House Speaker Nancy Pelosi slurring her words circulated widely on social media in Spring 2019 and gained more than 3 million impressions [11], the House Intelligence Committee noted the gravity of altered media and committee chairman Adam Schiff said that federal regulation of the technology is “worthy of serious consideration” [12]. (It is worth noting that the video of Speaker Pelosi would fall more into the “cheap fake” category rather than deepfake – see “Deepfakes Vs. Cheap Fakes” below.)

Taking a further step, in October 2019 California passed AB 730, which made illegal the “producing, distributing, publishing, or broadcasting campaign material, as defined, that contains either:

- a picture or photograph of a person or persons into which the image of a candidate for public office is superimposed
- a picture or photograph of a candidate for public office into which the image of another person or persons is superimposed, unless the campaign material contains a specified disclosure” [13].

In other words, sharing deepfakes is illegal in California, but only as it pertains to elections, and only within 60 days of an election, and only until 2023. However, civil rights groups were quick to condemn the law and question whether it would solve the underlying problem of deepfakes (see “What Are Our Options?” below) [14].

[12] https://apnews.com/4b1eac588bf5047a988b6f7ac4ac5ba7
Already, it’s clear that deepfakes’ use in U.S. elections is not a matter of “if” but “when.” In February 2020, just days before Legislative Assembly elections in Delhi, India, doctored videos of Bharatiya Janata Party (BJP) President Manoj Tiwari were circulated across 5,800 WhatsApp groups and reached roughly 15 million people [15]. This marked the first known use of a deepfake in a major political election. In this case, two deepfakes were reportedly produced and used as part of the campaign – that is to say, the group that produced the deepfakes said they were instructed to translate the videos into a new language and share the messages more widely. But regardless of whether the videos were created for “good or evil,” the fact that they were used at all shows that misinformation has taken a major step into the future. Tarunima Prabhakar, cofounder of Tattle, told Vice about the incident, “The problem with the ‘positive’ campaign pitch is that it puts the genie out of the bottle.”

At the same time, major national elections or country heads speaking about international issues are not the only potential feeding grounds for deepfakes. In parts of the world with less advanced information tracking, that have less media and press infrastructure, or that rely more on social media platforms to disseminate news, sharing even a basic deepfake among a smaller group of people could create major issues and even cost lives. Mounir Ibrahim is the Vice President Strategic Initiatives at Truepic, an image verification technology platform working on solutions to cheap fakes and deepfakes through the certification of images and videos. Ibrahim previously worked as a foreign service officer for the U.S. Department of State and shared his concerns about how deepfakes might be less problematic in terms of depicting someone like Kim Jong-Un falsely announcing a nuclear missile launch than of a smaller potential conflict brought about by misinformation in the form of a deepfake share on a social platform. “[What I’m more worried about] is some local tribal leader in a Sub-Saharan African country or in Asia or in India, maybe even parts of the United States, that sends something highly provocative, provoking immediate sectarian conflict or inter-communal conflict,” Ibrahim says [16].

Additionally, while, as of this article’s publication, there have been no reported incidents of deepfake use around the COVID-19 pandemic, the World Health Organization reported that alongside battling the actual disease itself, it was wrestling with what it dubbed an ‘infodemic,’ which it described as “an overabundance of information – some accurate and some not – that makes it hard for people to find trustworthy sources and reliable guidance when they need it” [17]. Already a situation fraught with concern, distrust, and panic, an incident such as the COVID-19 pandemic is ripe for the use of deepfakes to sow further uncertainty and drive rash decisions or greater complications.

DEEPPAKES IN BUSINESS

Similarly, we’ve yet to see any active use of deepfakes in the business sphere. From a still image perspective, companies in industries such as insurance, banking, and others that rely on factual data to process monetary claims are already deploying technology like Truepic to confirm that an image hasn’t been manipulated to, for example, exaggerate an insurance claim. Deepfake videos, however, may not be as much on the radar of business, but the possibilities are not difficult to imagine. As the fictional scenario at the beginning of the article demonstrated, someone could produce a deepfake with misinformation to manipulate a stock price and trade on the resulting swing. And as the India example shows, the information does not even need to be negative. A deepfake could portray a CEO of a biomedical company announcing the results of a major drug trial, leading to a jump in the stock price and an opportunity to trade on the swing. And again, the information does not need to stick for very long. Even if a company were to quickly respond and correct the record, just a few minutes of deepfake-produced misinformation could be enough for major stock swings and illegal stock trading opportunities. It may seem unlikely or even far-fetched, but unless businesses recognize this threat and prepare accordingly, only time will tell how bad actors might deploy these tools.

Furthermore, the general erosion of trust can have other implications in the business world. For several years (and especially following the 2016 U.S. Presidential elections), platforms such as Facebook and Twitter have come under fire for the ease at which they seem to help misinformation spread online. If left unattended, the increased sharing of deepfakes could very likely erode the general trust that users feel in what they see on these social networks. It might start at just a few videos, but if more and more content is either tagged as suspicious, unreliable, or outright fake, users may turn away from these channels as they see decreasing marginal utility as it becomes too difficult, or more likely too annoying, to separate fact from fiction. What happens to Facebook when people stop using it because they can’t believe anything on it? Perhaps a new business opportunity? (see “A Deepfake-Free Social Network?” below).
WHAT ARE OUR OPTIONS?

Whether approaching the problems of deepfakes as a business, politician, voter, or anyone else, it’s clear that as the technology evolves and more opportunities present themselves, the issues will only get worse. So, what should be done to fight back? As of now, a few options are on the table.

REGULATION/POLICING

The first option in addressing deepfakes would be regulation. As previously noted, California is the first U.S. state to pass any sort of legislation aimed at combating deepfakes. However, this legislation is focused on the time period shortly before elections and may not be far-reaching enough to tackle the problem. Furthermore, this legislation is targeted at punishing those who share the information, provided the perpetrator can be tracked, charged, tried, and found guilty – a tall order in a digital world. Additionally, we’ve seen that a deepfake’s damage can be done within just a few short hours, meaning that punishing violators would miss the actual crime itself and the fall-out caused. These complexities do not necessarily mean, that regulation is a dead-end. As with many other technologies, regulation is complex and tricky. However, perhaps the government could think more broadly about regulation outside of deepfakes themselves and more in terms regulating the tech platforms where the deepfakes are shared. While questions of big tech regulations are worthy of their own examinations in separate articles, it should be noted that the implications of deepfakes on platforms such as Facebook and Twitter bring into focus questions around how said services and platforms can be weaponized in certain situations. Professor Farid understands the complexity of these problems and how regulation can be complicated but nevertheless he feels that it is a necessary step. “We should just regulate [big tech platforms] in just the right amount of regulation, not too much, not too little,” he says. “But I don’t think this is going to come from inside [the companies]. It’s going to have to come from the legislation. And what's difficult here is, these aren’t really U.S. companies. These are global companies.” Given their reach, Professor Farid argues, it may fall on more holistic regulation of big tech companies overall, rather than deepfake-specific legislation, to address the problem. “We’ve allowed them to grow to a scale that is unprecedented and with remarkable power,” he says.
Stepping away from the federal regulation ledge, some of the larger tech platforms are taking it upon themselves to design and implement anti-deepfake policies. For its part, in January 2020, Facebook announced intentions to more aggressively fight against deepfakes in a blogpost titled “Enforcing Against Manipulated Media” [18]. According to these guidelines, Facebook commits to removing “misleading manipulated media” if it meets the following two criteria:

- “It has been edited or synthesized – beyond adjustments for clarity or quality – 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.
- “It is the product of artificial intelligence or machine learning that merges, replaces or superimposes content onto a video, making it appear to be authentic.

"The ban, however, does not include videos that could be argued to be “parody or satire” (a category of speech well protected by the First Amendment and tested by several court challenges [19]) or, as The Verge pointed out, “misleading edits made using traditional means, like [the] viral video of House Speaker Nancy Pelosi supposedly slurring her words” [20].

Others have further pointed out that this policy’s criteria for removing a deepfake – rather than flagging it as suspicious – only applies to videos that make a subject say words they didn’t actually say. Will Oremus of the tech publication OneZero noted that this loophole “would allow, for instance, a deepfake video that makes it look like a politician burned the American flag, participated in a white nationalist rally, or shook hands with a terrorist,” an example that a Facebook spokesperson confirmed would not be prohibited under this policy [21]. The policy also notes that independent third-party fact-checkers will also have opportunities to review videos flagged as potential deepfakes, though their rating could only lead to its being alerted in feeds as being potentially false, downgraded in the algorithm, and rejected as an advertisement. The deepfake would not be removed altogether. Facebook argued that “if we simply removed all manipulated videos flagged by fact-checkers as false, the videos would still be available elsewhere on the internet or social media ecosystem” – an unfortunate reality regarding the speed at which content can go viral not just on major platforms such as Facebook, but across the rest of the Internet as well.

[21] https://onezero.medium.com/facebooks-ban-on-deepfakes-is-a-half-step-at-best-99f602914db31
Other social media outlets have announced similar policies. In January 2020, Reddit update its policies to include language banning “deepfakes or other manipulated content presented to mislead, or falsely attributed to an individual or entity” [22] in an effort to not necessarily thwart existing threats but to “hedge against things that we haven’t seen much of to date, but could see in the future” [23]. In February 2020, Twitter also updated its rules regarding media that has been “significantly and deceptively altered or fabricated” with plans to label such tweets, warning people before they retweet or like the tweets, reduce visibility of the tweet, and provide additional context where possible [24]. And in April 2020, Google Play announced policies directed at removing apps in its Google Play app store that “promote or help create false or misleading information or claims conveyed through imagery, videos and/or text” [25].

To be sure, launching a new social network or platform is not an easy feat. Snapchat, launched in 2011, is the more recent major network – though TikTok’s recent rise could portend another major player in the space. By some counts, more than 60 social networks have launched and failed in the last decade, many of which tried to focus on certain content elements or focuses to differentiate themselves from platforms such as Facebook [28]. But perhaps the rise of deepfakes will generate enough of a burning platform for change that users will demand outlets that prioritize veracity of content. And if Facebook et al. aren’t able to move quickly enough and rethink their business models in a way that will suit these demands, new players could very well come along to seize on this opportunity.
TECHNICAL SOLUTIONS

While these policies are promising steps that not only recognize the apparent threat of deepfakes but take some steps towards mitigating their effect, they are mostly still targeted towards detecting or identifying deepfake media once it has already been posted, shared, and possibly even gone viral. From a technical standpoint, this is a very difficult problem. In 2019, Amazon Web Services, Facebook, Microsoft, and the Partnership on AI launched the Deepfake Detection Challenge inviting people to “build innovative new technologies that can help detect deepfakes and manipulated media” [26], but no real breakthroughs have been made. After all, as Professor Farid points out, “at the scale of 500 hours of YouTube video uploaded every minute, a billion uploads to Facebook a day, you can’t analyze your way out of this. You just can’t.” So, given that we are bombarded with thousands of pieces of digital media, images, videos, and audio clips every day, rather than trying to authenticate every single one, another option would be to put the burden of proof on the sender, the producer, or the recorder of the content. Using control capture technology, this line of thinking would require someone who records a video to use a certain app of type of app that would encode a set of data, hash it, cryptographically sign it, and then upload the necessary information into the blockchain. Professor Farid describes this process as a potential way to address these problems, saying this form of video capture “means at the point of recording you have this distinct signature that is now a distributed, immutable ledger of the blockchain.”

Professor Farid continues, “At any point in the future, somebody could send me [a file] and I can say, ‘Well, how do I know it’s authentic?’ To answer that, I could reach into the blockchain (or whatever your favorite mechanism is for storing that) and then authenticate it. In my view, that is the only technology that we have today, or in the near future, that will work at scale.” This is the type of technology that Ibrahim and Truepic have been working on. Using controlled capture technologies, Truepic is trying to establish “image provenance” by adopting a more forward-looking mindset in terms of how images are created going forward. As of now, most of Truepic’s business has been focused on controlled capture technology for still images, but video-based deepfakes are next on the horizon. One can imagine a future in which the onus falls less on verifying an image or video after the fact than on ensuring its creator has uploaded the necessary metadata for someone on the other end to verify the content’s veracity. This wouldn’t necessarily solve for recycling old imagery, but it could provide a blueprint for how to think about deepfakes going forward.

[26] https://deepfakedetectionchallenge.ai/
While tools like Truepic or other controlled capture technologies are promising, they are mostly still in the works and might not be widely available for quite some time. As of now, “there is no silver bullet for any individual to really detect it,” Ibrahim says. Instead, he argues, the best weapon might be, for the time being at least, awareness. “The most important thing is a level of education and resiliency to not overreact.” For businesses, this means a few things. Similar to how an organization might think about its cybersecurity regiment, businesses should educate all their employees and relevant constituencies about what deepfakes are and what the technology can do. Currently, public awareness of deepfakes is relatively low. But, like other “fake news” or misinformation online, it’s imperative to understand the threat and be vigilant in our daily lives. Unless enough people are aware of the potential of deepfakes, it will be very difficult to convince a critical mass of people that a video is adulterated if they didn’t even know it was possible.

Along the same lines, businesses can think about shoring up their overall public persona and communications strategy. Not all CEOs and executives need to be in the public eye, but most organizations would do well to build public trust in their company and its leadership so that if a deepfake does circulate, people will have an easier time identifying it as false if they are able to compare it to their mental image of a company and its CEO. If most people know your CEO and are used to hearing her say “X”, if a video circulates showing the CEO saying “not-X,” then people might be quicker to question the video as possible deepfake, or at least question the video in general and dig further for the truth. Already important, a robust and clear communications strategy will become increasingly critical for businesses thinking about fact versus fiction in this digital landscape.
Following general awareness, businesses should think about where a deepfake or other altered image might appear. Could a deepfake be used as part of a larger cybersecurity attack? Or simply to spread distrust among workers or between a union and management? What about targeted social media posts, or public appearances from high-level executives? The same way a business identifies its “crown jewels” in the process of building its cybersecurity systems, a company should examine its public image and communications channels to identify spots where a well-placed deepfake could do the most damage. And of course, having the team in place to properly maintain and monitor those channels is critical. Most businesses will already have people working on these channels (if the channel exists at all), but the added level of scrutiny is paramount. Every time your company name or an executive’s name is mentioned on social media, whether in text, inlaid in an image, or in a video, a vigilant business will be monitoring the content and confirming it is accurate and not a modification or adulteration. If not, it is imperative to move as quickly as possible to have the content removed, because the longer it’s live online, the greater potential it has to go viral and spread misinformation.

4 STEPS TO GUARD AGAINST DEEPFAKES

1. EDUCATE YOURSELF AND YOUR ORGANIZATION.
   You need to know that deepfakes are possible and what they can do.

2. SHORE UP YOUR OVERALL COMMUNICATIONS STRATEGY AND PUBLIC PERSONA,
   including any relevant executives. If people are used to hearing your CEO say “A” and a video circulates showing the CEO saying “NOT-A,” the public will hopefully be quicker to question the video as a possible deepfake.

3. IDENTIFY WHICH SPEAKERS, TYPES OF INFORMATION, OR FORUMS COULD BE MOST DAMAGED BY A DEEPFAKE
   Public appearances from high-level executives? Internal video content? Social media posts on specific topics?

4. MAINTAIN RECORDINGS OF EVERYTHING YOUR CEO, HIGH-LEVEL EXECUTIVES, OR OTHER KEY PEOPLE SAY.
   If a deepfake is released, you want to be able to counter it with the real video or imagery as fast as possible.
Lastly, this may be the most difficult, but vigilant organizations should proactively record everything the CEO says publicly. Professor Farid suggests this tactic in order to have a clear record in case it becomes necessary to dispute a deepfake and set the record straight. “Every time your CEO gets on stage, somebody is there recording with a control capture, so that if that thing ever gets modified, there’s no question as to what the original is,” he explains. Ideally, this would be done with controlled capture technology that can embed the necessary metadata that can be stored and checked against any altered imagery. But as a business waits for the necessary technology to become more widely available, maintaining its own database of verified content would be a strong first step to prepare for a nightmare scenario in which a deepfake goes viral with damaging misinformation.

These may seem like simple, or even simplistic, steps, but businesses are unfortunately in some ways at the mercy of public perception. Luckily, we’ve yet to see any cases of deepfakes being used to affect a company in one way or another. But with the technology rapidly improving, now is the time for businesses to prepare themselves for the threat of deepfakes. Going back to our initial scenario, the fact of the matter is that when these sorts of technologies are available, the safest option is to plan for the worst and hope for the best. Regulation is on the table, though whether that pertains to deepfakes specifically or social platforms in general remains to be seen, as does how effective regulation would be at managing this ever-changing issue. And technological solutions are on the horizon, but with the threats currently evolving faster than the solutions, waiting for a fix is not a sound business strategy. The good news is that public awareness and education are powerful tools and if enough of the general public can be alerted to the existence of deepfakes and on guard for them while browsing the Internet, we will hopefully be able to avoid the more disastrous outcomes and maintain a greater sense of digital trust.