Tim Jing

Dr. Kevin Moore

PWR 2KR

December 9th, 2023

Deepfakes and Propaganda

I. INTRODUCTION

On December 10th, 2016, Donald Trump posted the following infamous tweet to his account: "Reports by @CNN that I will be working on The Apprentice during my Presidency, even part time, are ridiculous & untrue - FAKE NEWS!" (Trump). The content is ridiculous, seemingly innocuous, and by that point, firmly integrated into Trump's oftentimes crass and faux-populist branding. However, this particular tweet is notable as it marks Trump's first usage of the term "fake news," which he would repeat thousands of times throughout his tenure as President.

In many ways, the propulsion of the term "fake news" into the mainstream lexicon during the precedent-shattering 2016 presidential election is symptomatic of our nation's gradual and continuous erosion of media trust. The country collectively dove headfirst into a media landscape riddled with landmines in 2016. PolitiFact's truth-o-meter swung progressively toward the left, blazing screens nationwide with "FALSE." And increasingly, Americans began doubting the information that crossed their devices. Whether it be parsing through deception wrapped in 99% truth or outright dismissing opposing viewpoints, "fake news" became an inevitable part of engaging with internet media—even the most well-respected institutions fell prey to the media hysteria. Text uncovered on the internet had become untrustworthy, and spin masters and grifters took advantage of that paranoia to sow discord and profiteer.

However, on the horizon today, perhaps more insidious and a more significant fundamental change than the shifting of text's role in society—which had always been used for propaganda purposes historically since the advent of the printing press—is the transformation of video and audio recordings. Deepfakes, a rapidly maturing technology where convincing audio and video can be synthesized with another person's voice or likeness, has largely flown under the radar compared to its potential to alter the fabric of society. Imagine a world where every security camera feed, "leaked" audio recording, and political speech can be and indeed needs to be questioned by fact-checkers for their authenticity. In today's modern age, virtually no one sees events in person and with their own two eyes—how can anything we see be truly trusted at that point? Indeed, video and audio have long enjoyed the epistemic authority that comes from their fundamentally immutable nature (Fallis 2021). Besides highly skilled doctoring using advanced, inaccessible technology which we will discuss at length, video and audio were trusted to maintain an accurate reflection of the world. Unlike text information that any anonymous user can imagine and spew out, video and audio needed to be based on reality. Destroying this exact epistemic authority is the terrifying potential of deepfakes, and it is critical that we as a collective stop and contemplate. We are at a potential inflection point. If this technology continues to proliferate and advance exponentially, we may enter a runaway scenario that we are unprepared for. Thus, the critical question is: are deepfakes an existential threat, merely an annoying yet omnipresent propaganda vessel, or something that fades into the background?

II. THE EARLY DAYS OF DEEPFAKES:

Deepfakes work through a simple premise. Machine learning (ML) algorithms, specifically neural networks, are trained on vast amounts of visual and audio data of people speaking. In a classic example, the ML model picks up on patterns—how mouth muscles move when a word is spoken, or the expressions in the eyes—and applies them to swap another person's face onto a pre-existing video (US Government Accountability Office). Similar manipulations can be done to alter speech or modify someone's facial expressions. The ML model, known as a Generative Adversarial Network (GAN), contains two parts: a generator and a discriminator. The generator tries to produce convincing, synthesized deepfakes, and the discriminator tries to determine whether something is fake. By receiving feedback from the discriminator, the generator begins identifying patterns to improve its deepfake generation capability, ultimately producing sometimes photorealistic replications that are indistinguishable to all but the most advanced discriminator algorithms (Kumar 2023).

The technical details aren't particularly important. What cannot be disputed is that the technology between "deepfakes" as we know it arrived in 2014, when famed researcher Ian Goodfellow first introduced the concept of a GAN ML model (Goodfellow 2014). However, it wasn't until an anonymous Reddit subreddit simply named "r/deepfakes" began circulating doctored explicit videos of celebrities in 2017 that the term garnered somewhat mainstream attention (Frąckiewicz 2023). Thus, deepfakes and the roots of crowdsourced propaganda are inextricably tied. Similarly, the term is also inextricably tied to sordid invasions of privacy and blackmail. Deepfakes create an intractable dilemma because, at its core, they close the "epistemic gap" between a random user and a public figure. In a pre-deepfake world, only the public figure had the authority to create a video or audio track featuring their likeness—any attempt at deception by a third party would be immediately recognizable. But suddenly, even as deepfakes weren't perfect, random individuals could create authentically-looking likenesses of public figures and post them on the internet. A person's features were no longer their own but instead "democratized"—and unfortunately, bad actors also receive unfettered access. Starting in 2018, researchers began developing countermeasures and ML tools to detect deepfaked technology. But we were complacent. Regulators, citizens, and even researchers sometimes felt that deepfake technology was immature. Most deepfakes were immediately recognizable, with telltale signs of manipulation and uncanny incongruence between speech and action. Subreddits like r/deepfakes could be shut down through a simple ban for involuntary pornography, with the mainstream audience blissfully ignorant of its existence. Thus, the threat wasn't brought up except among particularly concerned and specialized communities. However, the impact of our complacency was immediate, and the comeuppance was swift.

III. DEEPFAKES TODAY - CASE STUDIES

On February 24th, 2022, Russia invaded Ukraine. It was the largest military incursion into a European country since WWII, and the entire world was shocked. Ukrainians struggled to hold the lines as the ambush continued. With the power of hindsight, it's clear to all that the Ukrainians were never going to lay down and let the Russians take their territory. But on March 17th, just a few weeks after the invasion amid the initial chaos and worry, a video emerged on the internet—it was President Zelensky of Ukraine surrendering.

The video looks normal. President Zelensky isn't speaking in English, and so for many people worldwide, the meaning of the words is indecipherable. Astute observers might notice in the video that his lip movements are slightly off, but it's convincing enough. After all, the video is intentionally grainy to obscure potential artifacts left by the deepfake process. In addition, the video was strategically deployed to go viral in the Western non-Russian and non-Ukrainian-speaking world, as the language barrier provided a shroud of authenticity. There was no way to confirm whether or not the intonations, tone, and speaking manner sounded natural. If you didn't know it was ultimately a deepfake, you might have thought it was real (Allyn 2022). The video was hauntingly believable, and the stakes were desperately high.

This example is perhaps the most high-profile case of deepfakes infiltrating and potentially severely disrupting society. In the span of five short years, from 2017 to 2022, the explosion in artificial intelligence infrastructure and the amount of data available for training led to increasingly sophisticated deepfakes. Companies and governments with almost unlimited resources by modern standards could dedicate tremendous computational power to generate convincing deepfakes. Our tools, namely the detection algorithms and our eyes, were failing to keep up (Vaccari 2020).

However, this is only the most visible issue. After all, political or other high-profile targets of deepfakes receive the mainstream news coverage that we see every day. The problem of deepfakes certainly isn't constrained within this population of well-known individuals. As a matter of fact, I contend that deepfakes have the least significant effect on them, at least compared to what they currently experience. They are such authoritative, powerful, or important figures that other media manipulation tactics can be easily and justifiably deployed on them—for example, a sophisticated actor would likely have enough resources to create a virtual model of Zelensky and painstakingly frame-by-frame craft a fake video. Politicians and world leaders see such propaganda every day, and citizens are used to processing their fake or manipulated media.

Instead, deepfakes facilitate attacks from unsophisticated actors on everyday people. On social media, deepfake con artists prey on naive and perhaps more gullible individuals using celebrity likenesses. Be it children or the elderly, deepfaking beloved people like popular YouTuber Mr. Beast or nostalgic, inactive celebrities to ask for money is already extremely common. The internet provides both a veil of anonymity and kindling for a viral explosion, meaning even if only 1% of people fall for a deepfake, the monetary upside is massive. Other uses of deepfakes invade the privacy of everyday citizens. Because today's society is always connected to the internet, there's bound to be surprisingly detailed video and audio information containing your likeness. Be it through the aforementioned deepfaking of explicit videos or resurrecting someone back to life, no matter the intentions, the privacy of individuals is compromised. The impact is already devastating. Research estimates that anywhere between 96-99% of all deepfake videos on the internet are deepfake explicit videos, created and posted by anonymous users without the consent of any party depicted (Coleman).

IV. DEMOCRATIZING DECEPTION:

In order to evaluate the dangers posed by deepfakes to the average person, CGI and "movie magic" are commonly used frames of reference. After all, in the hands of a skilled visual effects (VFX) artist, fantastical scenes can be brought to life with tremendous fidelity and realism. In many instances, current CGI technology can surpass the accuracy of the average deepfake creation tool. So why hasn't CGI ignited the same controversy and discourse as deepfakes? I contend that deepfakes are a subset of CGI. After all, the Oxford English Dictionary defines "CGI" as "computer-generated imagery, special visual effects created using computer software, typically for use in film and television" (OED). Deepfakes, with its machine-learning architecture, ultimately accomplishes the same goal: generating imagery through computers that do not exist in real life. Interestingly, although they are so similar, deepfakes receive most of the attention. Perhaps this discrepancy in coverage is in and of itself propaganda—whether it is a coordinated fearmongering campaign, the compounding nature of an exciting news story, or truly something coincidental is unknown.

There is one key difference between the two technologies, however, that nonetheless renders deepfakes worth considering on its own as an emerging technique. VFX artists have undergone rigorous professional training to hone their skills. They are bound by a code of conduct that enforces strict standards in terms of professional conduct and the use of the technological prowess they possess (Visual Effects Society 2018). Furthermore, there are fundamentally fewer people who are skilled enough in traditional VFX or CGI to create believable and potentially damaging media. In contrast, deepfakes are accessible to everyone. While previous examples of deepfakes, like President

Zelensky’s surrender, are extremely realistic and require significant computational power, as alluded to above, basic deepfakes that can still fool the undiscerning eye are shockingly accessible. A cursory search on Google reveals numerous web apps—Deepfakes Web, Hoodem, Synthesia—that promise automatic generation of deepfakes with little work. For the more technically savvy, online open-sourced tools hosted on GitHub may be more powerful and nonetheless easy to use.

This democratization of CGI-like technology creates the threat we see today. At its core, deepfakes facilitate the impersonation of perspective. It’s a technique as old as deception itself—intentionally misattributing speech or information to others, be it to undermine their credibility or co-opt their authority to make a point, can be done by means outside of deepfakes. It can be as simple as slapping a “The Onion”" sticker on an otherwise non-ironic flyer or spreading misinformation online about a celebrity’s statements. However, deepfakes allow for perspective manipulation on a new level. Instead of manipulating fixed information, such as already-posted flyers or misappropriating an already-completed speech, deepfakes allow any user to hijack another person’s speech in real-time. A video, at least in today’s society, is relatively credible. If the video looks convincing, one generally assumes that the speaker did in fact at one point physically stand there, move their lips, and utter the words depicted.

Fortunately, although the dangers of such manipulation are profound, they aren’t all too unfamiliar either. An analogous form of propagandistic media would be out-of-context clipping of media. Similar to deepfakes, these clips intentionally misconstrue the speaker’s intentions and in fact possess more authority since they don’t involve any computer-generated edits to the video. In addition, they can also be circulated widely and virally. Politicians and other spinmasters engage in such behavior all the time. From mislabeling the dates and context of certain videos to deceptive editing, clips are constantly being shared on social media platforms that do not accurately reflect reality (The Washington Post 2019). Many times, these video clips are called out. However, the reality is that not everyone will see those corrections, and those who want to believe the video clip will believe it anyway. Confirmation bias and creating an echo chamber are extremely easy to fall into on the internet.

The impacts of this media manipulation, while scary, have become the status quo of today. It is already a given that media literacy and critical evaluation of sources shared online are required. Whenever we see a video clip, discerning viewers already question the source, editing, and context of the clip being shared. If that is the case, it is reasonable to assume that the status quo won't change as much with the more mainstream adoption of deepfakes. The redeeming quality of deepfakes is that subtle adjustments to already existing videos are the most deceptive—completely computer-generated videos are easier to detect with current detection tools. Thus, much like how a source video exists for each out-of-context clip, there will also be a source video to compare to for deepfaked videos. This way, how we evaluate media will not change significantly. Instead, deepfakes seem like they can just be folded into existing frameworks for video manipulation and recognition.

V. THE REAL THREAT OF DEEPFAKES: "VIRTUAL DOMINATION"

The more speculative and insidious danger associated with deepfakes comes from the damage to an individual's internal psyche from repetitive, targeted attacks using deepfake technology. Because of its democratized and computational nature, deepfakes can easily be deployed by individual actors to repeatedly harm a single person. Regina Rini, a philosopher and leading thinker on deepfakes, makes a similar comparison between CGI and Hollywood with deepfakes. While the ability of deepfakes to "provide an approximation of the same effect [as Hollywood], much more cheaply and quickly" is concerning to her, another critical aspect of deepfakes is that "Hollywood stars usually consent to being doppelgangered" (Rini 2022). In essence, Rini argues that this doppelganger duplication of a person's real-world likeness onto media like movies is consensual, and even if the doppelganger doesn't accurately reflect reality, there is no breach of contract. This distinction is critical, as this contractual awareness clearly defines the boundaries of what can or cannot be done through media manipulation technology. The actor or actress is well aware of what their impending virtual self will do, and when it will be broadcasted to the public. The movie company cannot abuse its position of power for fear of being sued. However, this distinction is lost with deepfakes. Deepfake artists can unleash their technology on anyone, and there is certainly no contract or legal recourse. A deepfake attack may indeed be illegal in the future—currently, only deepfakes that fall in line with existing criminal behavior like blackmail and involuntary pornography can be prosecuted—yet the anonymized nature of the internet makes low-profile deepfake attacks exceedingly resource-intensive to track down.

As described by Rini, the risk associated with this paradigm shift in media manipulation is that "deepfakes offer their creators a disturbing form of power over the people" it clones, which is no longer bound by mutual consent on both sides (Rini 2022). She coined the term "virtual domination" to refer to the power dynamic that occurs between the faker and the fakee, which is wildly lopsided, particularly since the attacker can hide behind a veil of internet anonymity. Rini specifically focuses on the pornographic industry with this term, which is warranted. This is indeed the most worrying application of deepfakes, as it is both the most prevalent and the most psychologically impactful to the victim. With deepfakes, a person's likeness can be "Frankenstein'ed." A face can be pulled from the victim, the torso can be of a celebrity, and limbs can be extracted from other people still. The body can be broken down and put back together using technology with no consent from anyone involved. This leads to hyper-objectification—in a world where people's bodies can be disassembled, the victim becomes more of a toy than a living, breathing human to the attacker. And clearly, this psychology is extremely dangerous. Simultaneously, virtual domination through deepfakes poses a potential loss of personal identity for the victim. Imagine constantly facing a barrage of disfigured, deformed visages of yourself engaging or speaking in ways you would never do. Imagine variations of the same video emerging a few times every day for a year on end. It would be debilitating and traumatizing. The daily act of denying these false statements would propel your fake videos into the limelight, yet you have to address them, as they can't simply be ignored. This catch-22 could create significant cognitive dissonance that damages the psyche. Unfortunately, with deepfakes, this could happen to anyone. A single person with a grudge and internet access could deliver a consistent payload of deepfaked media. This is what virtual domination would look like, and it's certainly within the realm of possibility for the near future. We've already seen cyberbullying, harassment, revenge porn, and other forms of internet blackmail cause psychological damage and direct loss of life. Deepfakes will undoubtedly exacerbate the dangers of targeted cyber attacks.

VI. THE POSITIVE APPLICATIONS:

While deepfakes today are traditionally associated with propaganda and misleading information, as an emerging technology, it is also imperative to consider different applications of deepfakes that may emerge in the future. Besides simply gaining greater fidelity and accuracy on lower and lower computational power (which is a given considering the rapid pace of technological innovation in the past few years), deepfakes may also populate several unique and creative niches that can contribute positively to society.

The first and most obvious one would be entertainment. The aforementioned connection between deepfakes and "movie magic" is already extremely clear. As the technology matures, beloved actors can turn back time to their primes or even be resurrected from the dead to star in movies or TV shows. If done with explicit permission and care, this could be a net boon for fans and the families of the celebrities. However, there is the omnipresent danger of privacy invasion.

Another sector ripe for innovation would be education. From bringing historical figures to life to generating visualizations of complex biological processes, deepfake technology could be leveraged to drive engagement. Here, the benefit of accessibility is demonstrated. With traditional CGI or visual effects, so much expertise is necessary that everyday teachers and schools are unable to access the technology. However, by democratizing this tool, there will also inevitably be creative users who will leverage deepfakes to improve more niche subsectors like creating educational virtual assistants that wouldn't traditionally be profitable for another technology. Finally, deepfakes could be a tremendous boon for influencer marketing. Specifically, influencers like college athletes or traditional celebrities would be able to create more promotional material with less time investment by licensing out their likenesses. As long as the system for verification of authenticity and licensing is robust, this could open up more opportunities for creative marketing.

VII. CONCLUSION AND RECOMMENDATIONS:

Ultimately, deepfakes, as a sub-branch of computer-generated graphics, possess a single defining characteristic that separates them from other forms of media manipulation: deepfakes are extremely accessible yet allow for a relatively accurate generation of false information. In that sense, the dangers of deepfakes arise from crowdsourced and grassroots use. While the dangers of deepfakes for propagandistic purposes might be overstated—after all, VFX and out-of-context clipping already simulate much of what deepfakes can do today, not to mention the perspective impersonation that takes place outside of video media—deepfakes remain a threat because it can be weaponized to target individuals.

Seeing deepfaked videos throughout social media and having to critically evaluate the authenticity of the video is one thing. It is an entirely different situation when someone engages in what Rini calls "virtual domination," where thousands of deepfaked videos of you specifically are spread throughout the internet, perhaps with severe invasions of privacy involved. Thus, deepfakes do indeed pose a threat. There must be extremely active regulation and control over criminalizing the use of someone's likeness without their permission. We must not focus solely on detecting deepfakes because it isn't the biggest threat. A video that most, if not everyone, would recognize as a deepfake could still be damaging and hurtful to the psyche.

Addressing the problem of deepfake technology and mass media manipulation is exceedingly difficult. A two-pronged response, both from the regulatory agencies that prosecute these virtual crimes and from the leaders of the technology platforms where media spreads is necessary, not to mention the continual engagement citizens must partake in to raise awareness. One particular solution is promising: I envision a future where, much like how intellectual property operates currently, a person's likeness belongs to themselves and is something that they will be able to license out. We already protect intangible concepts like branding and various inventions with trademarks and patents—what is so different about a person's visage that shouldn't be protected? Perhaps as part of this policy, the government will institute a digital signature that can be incorporated into online media to prove the authenticity and official licensing of an individual's likeness. The digital watermark, which doesn't purely have to be visual and can in fact be embedded into the metadata of the video itself, could be distributed by the likeness license holder and automatically detected by video players. Without the digital signature, deepfaking someone else would be akin to violating a patent today. After all, we all deserve our privacy. We deserve to know and control where and how our faces emerge in the media, deepfaked or not.

References

Allyn, Bobby. 2022. "Deepfake Video of Zelenskyy Could Be 'Tip of the Iceberg' in Info War, Experts Warn." *NPR*, March 16, 2022, sec. Technology. https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia.

Coleman, Kate. n.d. "How Deepfakes Are Impacting Culture, Privacy, and Reputation." Statuslabs.com. https://statuslabs.com/blog/what-is-a-deepfake.

"Deconstructing Deepfakes—How Do They Work and What Are the Risks? | U.S. GAO." 2020. Www.gao.gov. US Government Accountability Office. October 20, 2020. https://www.gao.gov/blog/deconstructing-deepfakes-how-do-they-work-and-what-are-risks#:~:text=Deepfakes%20rely%20on%20artificial%20neural.

Fallis, Don. 2020. "The Epistemic Threat of Deepfakes." *Philosophy & Technology* 34 (August). https://doi.org/10.1007/s13347-020-00419-2.

Frąckiewicz, Marcin. 2023. "The Evolution of Deepfake: Tracing the History and Development of AI-Generated Content." TS2 SPACE. July 19, 2023. https://ts2.space/en/the-evolution-of-deepfake-tracing-the-history-and-development-of-ai-generated-content/.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. "Generative Adversarial Nets." Neural Information Processing Systems. Curran Associates, Inc. 2014. https://papers.nips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

Kumar, Nitesh. 2023. "What Is Deepfake Technology? Origin and Impact." Analytics Insight. June 25, 2023.

https://www.analyticsinsight.net/what-is-deepfake-technology-origin-and-impact/#:~:text=Origin%20of%20Deepfake%20Technology%3A.

Rini, Regina, and Leah Cohen. 2022. "Deepfakes, Deep Harms." *Journal of Ethics and Social Philosophy* 22 (2). https://doi.org/10.26556/jesp.v22i2.1628.

The Washington Post. 2019. "The Washington Post's Guide to Manipulated Video," June 25, 2019. https://www.washingtonpost.com/graphics/2019/politics/fact-checker/manipulated-video-guide/.

Trump, Donald. 2016. "Https://Twitter.com/RealDonaldTrump/Status/807588632877998081?Lang=En." X (Formerly Twitter). December 10, 2016. https://twitter.com/realDonaldTrump/status/807588632877998081?lang=en.

Vaccari, Cristian, and Andrew Chadwick. 2020. "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News." *Social Media + Society* 6 (1). https://doi.org/10.1177/2056305120903408.

Visual Effects Society. 2018. "Visual Effects Society Code of Conduct." Visual Effects Society. May 9, 2018. https://www.vesglobal.org/visual-effects-society-code-of-conduct/.