Are Deep Fakes a Shallow Concern? A Critical Analysis of the Likely Societal Reaction to Deep Fakes

Jeffrey Westling[1], jwestling@rstreet.org

Abstract

Deep fakes, a class of AI generated audio-visual materials designed to appear as an authentic record of actual speech, garnered increasing attention as worries about foreign disinformation campaigns and so called "fake news" have increased. Some critics have raised concerns that this new technology is too realistic for viewers to differentiate fact from fiction, which allows bad actors to manipulate elections, induce societal unrest, and incite panic. From their perspective, the influx of deep fake content may lead to the death of trust in media outright, as people will assume that all content may be artificially generated "fake news."

Yet close consideration of such an argument reveals a key assumption upon which it is based: namely, that deep fakes, once they are technologically advanced and easy to produce, will either be believed without question or will fundamentally shift public perceptions of video such that even real ones will be dismissed. However, these are not the only two possible outcomes and thus the purpose of this paper is to present a more nuanced view.

Indeed, as is so often the case, the likely societal reaction to deep fakes will likely fall somewhere between these two extremes, wherein society develops proxy mechanisms for assessing the reliability of video evidence in the wake of deep fake technology. Such a likelihood is based on two classes of observations. First, the reason we trust images and video may stem largely from societal norms about the use of the particular medium, rather than something inherent to the medium itself. Second, history has shown that similar concerns about digital photo editing techniques did not lead to either of the outcomes predicted for societal perceptions of truth, and thus how society reacted to fake photos sheds much light on what is likely to happen with regard to deep fake videos.

Identifying the likely and less-drastic social trends in reaction to deep fakes is exceptionally important today, because ongoing fears of the technology have prompted calls for regulatory responses. For example, some have proposed amending Section 230 of the Communications Decency Act ("CDA 230") to increase liability on platforms that do not take reasonable steps to limit the spread of deep fake content. To the extent that there is a likely scenario in which the ordinary operations of society adapt to manage the impact of deep fakes on perceptions of truth, the need for policy responses (that no doubt will be imperfect and potentially detrimental to valuable technological advances) is strongly reduced.

This paper proceeds by covering two main areas: emerging technologies and online platform regulation. It first explains the likely reaction to deep fakes by reviewing the development of similar technologies as well as key distinctions from these technologies of the past. Second, the paper examines whether regulatory responses to deep fakes, focusing primarily on the calls to amend CDA 230, are necessary or whether existing regulatory tools and free market forces will be sufficient.

[1] I would like to thank Blake Reid and Jack Karsten for their insightful comments on the draft.

Contents

I. Deep Fakes: What are They and What do they Do 2
a. The Nature and Context of AI-Generated "Deep Fakes" 2
b. Uses for Deep Fake Technology 3
II. Deep Fakes in Context: Historical Review of Image Manipulation and Societal Response 5
a. "Seeing is Believing" 6
b. The Rise of Digital Manipulations and Industry Responses 12
III. Key Distinctions to Consider 15
a. Social Media 15
b. Accessibility of the Technology 17
IV. Societal Reactions to Deep Fakes 18
a. Expected Societal Responses 18
b. Regulatory Approaches 22
Conclusion 24
About the author 25

Any time a new technology enters the marketplace, society often considers worst case scenarios as an inevitability[2] and so far, this has proven no different in the case of artificially generated synthetic media, commonly known as deep fakes.[3] However, as has also been the case in the past, although harms may occur, the impact of the new technology itself—will very likely pale in comparison to the worst case fears.

This paper examines society's likely response to deep fakes in the context of a similar historical development: the development of digital photographic editing tools. While some may have claimed that digital photographic editing would lead to the death of truth in the medium, photographs always lied. Institutions and artists that used the medium to depict the world as it occurs drove trust in the medium, and the doomsday scenarios some feared would arise as the technology advanced never came to pass. Deep fakes will likely go a similar route, with some harms occurring, but the scenario where we simultaneously believe everything and nothing will not materialize.

First, the paper explains the technology that generates deep fake media and the uses, both good and bad, for the technology. Second, it looks to put deep fake in context, showing that society has never relied solely on the content itself as a source of truth. Third, the paper introduces key distinctions to consider, but ultimately explains that these distinctions do not present a unique issue for deep fake media. Finally, the paper examines society's likely response to the technology, as well as potential regulatory responses that could supplement a hands off approach. The main concerns with deep fakes stem from a fear of two equally worrisome but contradictory ideas: deep fakes will deceive everyone or cause them to reject all

[2] Adam Thierer, "Technopanics, Threat Inflation, and the Dangers of an Information Technology Precautionary Principle," Minn. J. L. Sci. & Tech. 14:1, 2019 p. 309.
[3] For example, Senator Marco Rubio (R-FL) argued that "all you need is the ability to produce a realistic fake video that could undermine our elections" to threaten the United States. Derek B. Johnson, "Rubio warns on "deep fakes" in disinformation campaigns," *FCW*, July 16, 2018. https://bit.ly/2ZsFR8L.

video as fake. These fears, while important to consider, fail to recognize that a third, more neutral path in which society adapts to the technology.

I. DEEP FAKES: WHAT ARE THEY AND WHAT DO THEY DO

a. The Nature and Context of AI-Generated "Deep Fakes"

Deep fakes, as the term is used popularly and in this paper, are a class of simulated audiovisual materials designed to appear realistic.[4] This new technology uses generative adversarial networks (GANs),[5] which incorporate both a generative model and discriminative model to simultaneously generate and detect synthetically generated content:[6]

> *The generative model can be thought of as analogous to a team of counterfeiters, trying to produce fake currency and use it without detection, while the discriminative model is analogous to the police, trying to detect the counterfeit currency. Competition in this game drives both teams to improve their methods until the counterfeits are indistinguishable from the genuine articles.*[7]

Essentially, the generative algorithm, creates new data instances while the second, the discriminative model, attempts to classify input data and predict a label or category to which the data belongs.[8] Because the Network already knows which images are "fake" and which are "authentic", the discriminator net can "learn" why the fake was or was not detected. However, as the discriminatory net improves its ability to label the generated inputs, the generator also improves its outputs because it too is "learning" what does, and more importantly, what does not get past the discriminator.[9]

As should be apparent from the description above, there is nothing about GANs that is specific to fake video-making; it is a general-purpose technology adapted to all sorts of tasks. In fact, the term deep fake itself refers to a variety of different models for creating synthetic media.[10]

In that sense, they fall within a long line of deceptive audiovisual practices that have been used since time immemorial. As far back as the 6th century BC, Sun Tzu said in the *Art of War*, "[a]ll warfare is based on deception."[11] Hannibal set up a fake camp to deceive pursuing Romans, leading them into an ambush.[12] Relatively more recently, Will Rogers impersonated President Harry Truman, fooling enough of his audience that the White House had to release a statement addressing the issue.[13]

---

[4] *See* Robert Chesney & Danielle Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," California Law Review 107 (forthcoming), 2019. https://bit.ly/2C78jnQ.
[5] Ian Goodfellow et al., "Generative Adversarial Networks," June 2014. https://arxiv.org/abs/1406.2661.
[6] Ibid at p. 1.
[7] Ibid.
[8] In the context of artificial media, the discriminator categorizes images as either authentic or generated. The network feeds images both from an existing data set and the generated images from the generative net. Ian Goodfellow et al., "Generative Adversarial Networks," June 2014, p. 1. https://arxiv.org/abs/1406.2661.
[9] Ibid.
[10] These techniques include individualized simulated audio, digital editing tools, facial reenactment, facial reconstruction and lip sync, and the swapping of specific body parts onto another individual. "Mal-uses of AI-generated Synthetic Media and Deepfakes: Pragmatic Solutions Discovery Convening," *Witness*, June 11, 2018. https://bit.ly/2nae100.
[11] Sun Tzu, The Art of War, I. 18. https://goo.gl/R0mKY.
[12] Livy, *Ab urbe condita*, 22.4.
[13] Elizabeth Blair, "Impersonating the President: From Will Rogers to Obama's Anger Translator," *NPR*, Oct. 29, 2012. https://goo.gl/6fA3cc.

Even image manipulation techniques are not especially new. William H. Mumler was producing fraudulent "spirit photographs" as early as the 1860s,[14] and Stalin famously spliced out individuals from photos after they had fallen out of favor with his regime.[15] Digital graphics technology presented new opportunities for forgers to create more realistic images, sounds, and videos.[16] And commercially successful applications such as Adobe Photoshop allowed anyone with a computer to produce manipulated images that could fool most viewers.[17]

Deep fakes are thus not a remarkable disruption or technological breakthrough in themselves. Rather, they represent a fairly predictable advance in a long tradition of deceptive practices, simply incorporating the latest in general-purpose technologies.

b. Uses for Deep Fake Technology

This automated technology for making simulated-reality videos can serve a wide variety of legitimate purposes, such as art,[18] education,[19] and even missing persons investigations.[20]

In education, for example, teachers can show their students videos of historic figures talking directly to the camera.[21] For students that may not find the subject very interesting, a realistic image depicting the subject of the lessons can make history come to life, allowing them to engage and connect with the lessons of the day.

Likewise, artists can create better works for the public. It is no surprise that filmmakers have been using digital editing to make their subjects appear differently than they could possibly in real life. In *Tron Legacy*, a video f/x team used computer-generated imagery to make it appear as though Jeff Bridges was back in his youth, allowing them to create the main antagonist of the film.[22] More recently, *Star Wars Rogue One* shocked audiences with a realistic recreation of a young Princess Leia.[23] The scene allowed the audience to connect emotionally and reminisce about the original release of *A New Hope* almost 40 years earlier. With deep fake technology, filmmakers can create similar scenes without the major financial investment needed to achieve the results with traditional CG methods.

---

[14] Peter Manseau, "Meet Mr. Mumler, the Man Who "Captured" Lincoln's Ghost on Camera," *Smithsonian Magazine*, Oct. 10, 2017. https://bit.ly/32xExnM.
[15] Erin Blakemore, "How Photo's Became a Weapon in Stalin's Great Purge," *History*, Apr. 20, 2018. https://bit.ly/2rru4sk.
[16] Hany Farid, "Image Forgery Detection: A Survey," IEEE Signal Processing Magazine, 2009, p.1. https://goo.gl/jkMvGB.
[17] See e.g. David S. Waller, "Photoshop and Deceptive Advertising: An Analysis of Blog Comments," *Studies in Media Communications* 3:1, 2015, https://bit.ly/2XHDG50; Matthew Gault, "Fake Images of Parkland Survivors Tearing Up the Constitution Go Viral," *Motherboard*, Mar. 26, 2018, https://goo.gl/tS8sQ8.
[18] *See, e.g.*, Eating Things, "Breaking: Jon Snow Finally Apologized for Season 8," *Youtube*, June 13, 2019. https://bit.ly/2X8Ne8b.
[19] Robert Chesney & Danielle Citron, "Deep Fakes: A Looming Crisis for National Security. Democracy and Privacy?" *Lawfare*, Feb. 21, 2018. https://bit.ly/2EP4nvf.
[20] Grigory Antipov et al., "Face Aging With Conditional Generative Adversarial Networks," Feb. 7, 2017. https://arxiv.org/abs/1702.01983.
[21] Robert Chesney & Danielle Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *California Law Review* 107 (forthcoming), 2018, pp. 14-15. https://bit.ly/2C78jnQ.
[22] Ryan Nakashima, "'Tron: Legacy' reverse-ages Jeffrey Bridges," *Today*, Dec. 7, 2010. https://goo.gl/iNWdUt.
[23] Matt Miller, "This Actress Secretly Played Princess Leia in Rogue One," *Esquire*, Mar. 14, 2017. https://goo.gl/QnTSZB.

The same technology can also be used to apply age progression models to a photo of an individual and generate an image that ages the subject.[24] And, in addition to its artistic uses, it can also be used by law enforcement to help generate images of children who went missing years ago.

Even outside of the context of deep fake images and videos, GANs present numerous opportunities for other media functions. For example, they have been used as a means of synthesizing images from text descriptions.[25] Such a function can make finding an image or gif on the internet much simpler or help readers visualize text in their favorite novels.

Put simply, despite the bad reputation that pervades the media, there are countless ways this technology can provide benefits to consumers. Many of these ideas are still in their infancy, and many more have yet to be even conceived.

c. Contemporary Concerns

While deep fakes present significant new opportunities, due in large part to the 2016 election, many argue that their risks outweigh their benefits.[26] And these worries are exacerbated by the fact that the technology to create deep fakes is becoming increasingly ubiquitous. While fake videos were previously limited to expert forgers and master tacticians, amateurs can now use fairly common and widely accessible AI technology to generate realistic HD video, audio, and document forgeries at scale.[27] And, given how little time it takes to click and share one of these videos once posted, deep fakes have the potential to spread at alarming rates and thus to become increasingly pervasive in modern society.

In large part it is for these reasons that the immediate reaction of most commentators has been that deep fakes are a categorically worse form of deception than digital photographic editing, or even image based memes and "fake news" stories.[28] As the argument goes, deep fakes, as audiovisual media, can resonate with a variety of different senses all at once. The audience hears the voice and sees the image moving, making the video all the more convincing.

Some commenters worry that the existence of widespread AI forgery capabilities will erode social trust.[29]

> *We will struggle to know what to trust. Using cryptography and secure communication channels, it may still be possible to, in some circumstances, prove the authenticity of evidence. But, the "seeing is believing" aspect of evidence that*

[24] Zhifei Zhang et al., "Age Progression/Regression by Conditional Adversarial Autoencoder," Mar. 28, 2017. https://arxiv.org/pdf/1702.08423.pdf.
[25] Ayushman Dash et al., "TAC-GAN – Text Conditioned Auxiliary Classifier Generative Adversarial Network" Mar. 26, 2017. https://goo.gl/Eb6JnC; Han Zhang, "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks," Aug. 5, 2017. https://goo.gl/YFKEBQ; Scott Reed et al., Generative Adversarial Text to Image Synthesis," June 5, 2016. https://goo.gl/MCpKHW.
[26] Robert Chesney & Danielle Citron, "Deep Fakes: A Looming Crisis for National Security, Democracy and Privacy," *Lawfare*, Feb. 21, 2018. https://goo.gl/YpDGFt.
[27] Ibid.
[28] *See* Joe Andrews, "Fake news is real — A.I. is going to make it much worse," USA Today, July 12, 2019. https://bit.ly/2SsIbux; "Ahead of 2020, Beware the Deepfake," *The Atlantic*, July 02, 2019. https://bit.ly/2xK9Nlj; Lisa Eadicco, "There's a terrifying trend on the internet that could be used to ruin your reputation, and no one knows how to stop it," *Business Insider*, July 10, 2019. https://bit.ly/2ShNMDL.
[29] Greg Allen & Taniel Chan, "Artificial Intelligence and National Security," *Belfer Center Study*, July 2017. https://goo.gl/D9n1FX.

> *dominates today—one where the human eye or ear is almost always good enough—will be compromised.*[30]

Such hypotheticals about deep fakes taken to the extreme present a bleak picture of the future. For example, fake videos could feature a public official taking bribes, displaying racism, or engaging in adultery immediately before an election in an attempt to swing voters.[31] Falsified video appearing to show a Muslim man celebrating the Islamic state could lead to violence against that community.[32] Or perhaps a fake video might depict emergency officials "announcing" an impending missile strike in a major metropolitan area.[33]

What is potentially worse is that even if one can trust a specific source of news, that news source may itself be fooled by the technology and accidently publish a fake video. Without a technical way of authenticating the video, even a reputable news source that does their due diligence may be unable to differentiate fake videos from authentic ones. When the news can no longer be trusted, people may retreat deeper into unfounded belief and all video and audio evidence might simply be dismissed as fake news.[34] However, this trajectory of events is predicated on the assumption that people automatically trust what they see in a video regardless of its context. And this simply may not be the case.

II. DEEP FAKES IN CONTEXT: HISTORICAL REVIEW OF IMAGE MANIPULATION AND SOCIETAL RESPONSE

In 1990, the year Adobe first released the commercial version of Photoshop, *Newsweek* published an article entitled "When Photographs Lie,"[35] which argued that the potential consequences of the rise in photographic manipulation techniques could be disastrous: "Take China's leaders, who last year tried to bar photographers from exposing their lies about the Beijing massacre. In the future, the Chinese or others with something to hide wouldn't even worry about photographers."[36]

These concerns were not entirely without merit. Fred Ritchin, who was formerly the *New York Times Magazine* picture editor and is currently the Dean Emeritus of the International Center of Photography School, has continued to argue that trust in photography has eroded over the past few decades:

> *There used to be a time when one could show people a photograph and the image would have the weight of evidence—the "camera never lies." Certainly photography always lied, but as a quotation from appearances it was something viewers counted on to reveal certain truths. The photographer's role was pivotal, but constricted: for decades the mechanics of the photographic process were generally considered a guarantee of credibility more reliable than the photographer's own authorship. But this is no longer the case.*[37]

Certainly, as Ritchin notes, the camera can and often does lie when the final product has been manipulated. But this is nothing new, and since the outset of photography this manipulation was

---

[30] Ibid at p. 31.
[31] Robert Chesney & Danielle Citron, "*Deep Fakes: A Looming Crisis for National Security, Democracy and Privacy,*" *Lawfare*, Feb. 21, 2018, https://goo.gl/YpDGFt.
[32] Ibid.
[33] Ibid.
[34] Ibid.
[35] "When Photographs Lie," *Newsweek*, July 29, 1990. https://bit.ly/2Xt8dPK.
[36] Ibid.
[37] Fred Ritchin, "What a Photograph Can Accomplish: Bending the Frame by Fred Ritchin," *Time*, May 29, 2013. https://goo.gl/S2K5bM.

expected. And now, with similar claims being made about deep fake media, it is important to examine the history of photographs and how society reacted to the development of a similar, truth altering technology. As it turns out, it may not be the content itself that drives trust in media but rather the surrounding context and institutions that disseminate and share the content.

a. "Seeing is Believing"

Many implicitly understand photographs, and subsequently video, to accurately depict real events because we can "believe what we see." With deep fakes, then, this new medium for creating fake videos could shock society's foundation to its very core, causing distrust in things we once thought credible. But perhaps these new developments may not cause this harm because the fears are based on a faulty premise. It has never been the case that simply seeing something out of context could provide a full and accurate picture of the "truth," an accurate saying, and instead a significant reason we trust images and video was the development of societal norms surrounding the use of photography as a tool for conveying ideas. And because of this, as photographic editing techniques and technology advanced, society didn't see the total breakdown in trust that some feared. Perhaps the harms deep fakes will cause are likewise limited by the simple fact that these societal norms can likewise be leveraged or changed as deep fakes become more common.

To understand the impact that this could have, it is first important to put "seeing is believing" in context. During photography's infancy, people viewed the practice more as an artform with the final product a scene the photographer created.[38] These images didn't serve as a source of undeniable truth, but rather as a single and particular commentary on their subject.[39] As such, viewers likely understood that the image was just that: art. Manipulation and staging were expected.

For example, in 1846, Calvert Jones' captured the photograph "Capuchin Friars, Valetta, Malta" which depicted four monks in Malta.[40] Jones took the image to serve as a pseudo-postcard for audiences on vacation to the Mediterranean Island to send friends and families back home.[41] It simply served as a fun way to keep in touch and share a slice of life. Being such an early image, it may seem unlikely that the photo could be drastically manipulated. Nor is there a strong incentive to manipulate such an image.

And yet it was manipulated. In fact, Jones cut out a fifth friar from the photo entirely by blacking out the individual on the negative image.[42] While a rudimentary technique, this was relatively common practice.

[38] Mia Fineman, FAKING IT: MANIPULATED PHOTOGRAPHY BEFORE PHOTOSHOP (The Metropolitan Museum of Art, 2013), p. 45 ("FAKING IT").
[39] Ibid.
[40] Calveret Richard Jones, "Capuchian Friars, Valetta, Malta, 1846" *The Metropolitan Museum of Art*, last visited July 24, 2019. https://bit.ly/320qCpK.
[41] FAKING IT, p. 4.
[42] Ibid, pp. 3-4.

Figure 1: A comparison of "Capuchin Friars" and the negative used to create it. The fifth monk is visible, but darkened out, in the negative, allowing the final to appear as though only four monks are present.[43]

Even images that conform with a contemporary understanding of "memes" were quite common. Some photography concessions at tourist attractions such as amusement parks and air shows would sell novelty images,[44] such as the guests flying a plane or riding a giant butterfly.[45] The most popular example was known as "horsemaning," in which it appeared as though the subject had lost his or her head.[46]

[43] Calvert Richard Jones, "Capuchian Friars, Valetta, Malta, 1846" The Metropolitan Museum of Art, last visited June 28, 2019. https://bit.ly/320qCpK.
[44] Ibid, p. 124.
[45] Ibid.
[46] Ibid, p. 117.

Figure 2: An individual juggling his own head.[47]

All of this is to say that, as Mia Fineman, Assistant Curator at the Metropolitan Museum of Art explained:

> [T]hat's what 19th century viewers wanted. They, 19th Century viewers were, especially in the late part of the century, were actually fairly accepting of photographic manipulation. They didn't always expect photographs to be eyewitness accounts of something that really happened. Photographs were seen as a kind of picture, similar to pictures

[47] Unidentified artist. Published by Allain de Torbéchet et Cie. *Man Juggling His Own Head*, ca. 1880. Collection of Christophe Goeury. Retrieved from "Faking It: Manipulated Photography Before Photoshop," The Metropolitan Museum of Art, 2013. https://bit.ly/2IK4KVR.

> like prints or drawings. It was really part of the graphic arts. It didn't always have to represent the absolute truth.[48]

But this raises an interesting question: why do we now adhere to the adage that seeing is believing if photography, and visual representations generally, were constantly manipulated and understood as such? As with most art forms, the use of the medium, as did society's understanding of it, changed over time. Part of this can be traced back to artistic evolution, with many viewing photography as a way to promote (or reinforce would work too as you wish) realism as an art form. However, what appears to have shifted this change in understanding about photography were the societal norms surrounding the use of photographs that developed during the rise of photojournalism.

During the late-nineteenth and early-twentieth centuries, photographs did not commonly appear in the news or as a record of events.[49] One reason, mainly in the context of news reporting, was a stigma that accompanied the use of images to convey information.[50] Many saw the practice of illustrating news as catering to the lower class, with at least one individual going so far as to argue that the illustrations contributed to illiteracy.[51]

A more significant factor, however, was simply the limitations of the technology as a means of capturing an authentic record of events.[52] The equipment was cumbersome and expensive, and there were no quick ways of capturing images and getting them to press.[53] Some used photographs to spread disinformation, but the inherent limitations meant that these campaigns were much more difficult.

For example, after a disastrous war against Prussia, the French Assembly declared itself the Third Republic.[54] This development worried Parisians who, in turn, declared their own government based in Paris.[55] During the subsequent fighting, the Parisian Communards threw up barricades, shot hostages, and burned numerous historical buildings in the city.[56] Roughly six months after the French Assembly's victory, Ernest Eugene Appert released "Crimes de la Commune," which was comprised of nine photographs depicting the insurrectionists' brutality. While the images were based on real events and captioned accordingly, Appert fabricated the images by hiring actors and staging the scenes, pasting the headshots of key participants atop the actors' bodies.[57]

---

[48] Press Event, "Faking It: Manipulated Photography before Photoshop," National Gallery of Art, July 1, 2013. https://bit.ly/2YmIjRp.
[49] Claude Cookman, AMERICAN PHOTOJOURNALISM: MOTIVATIONS AND MEANINGS (Northwestern University Press, 2009), p. 66 ("AMERICAN PHOTOJOURNALISM").
[50] Ibid.
[51] Ibid.
[52] Ibid, pp. 144-45.
[53] Ibid.
[54] Adam Gopnik, "The Fires of Paris," *The New Yorker*, Dec. 15, 2014. https://bit.ly/2SAgpv3.
[55] Ibid.
[56] FAKING IT, p. 95.
[57] Ibid.

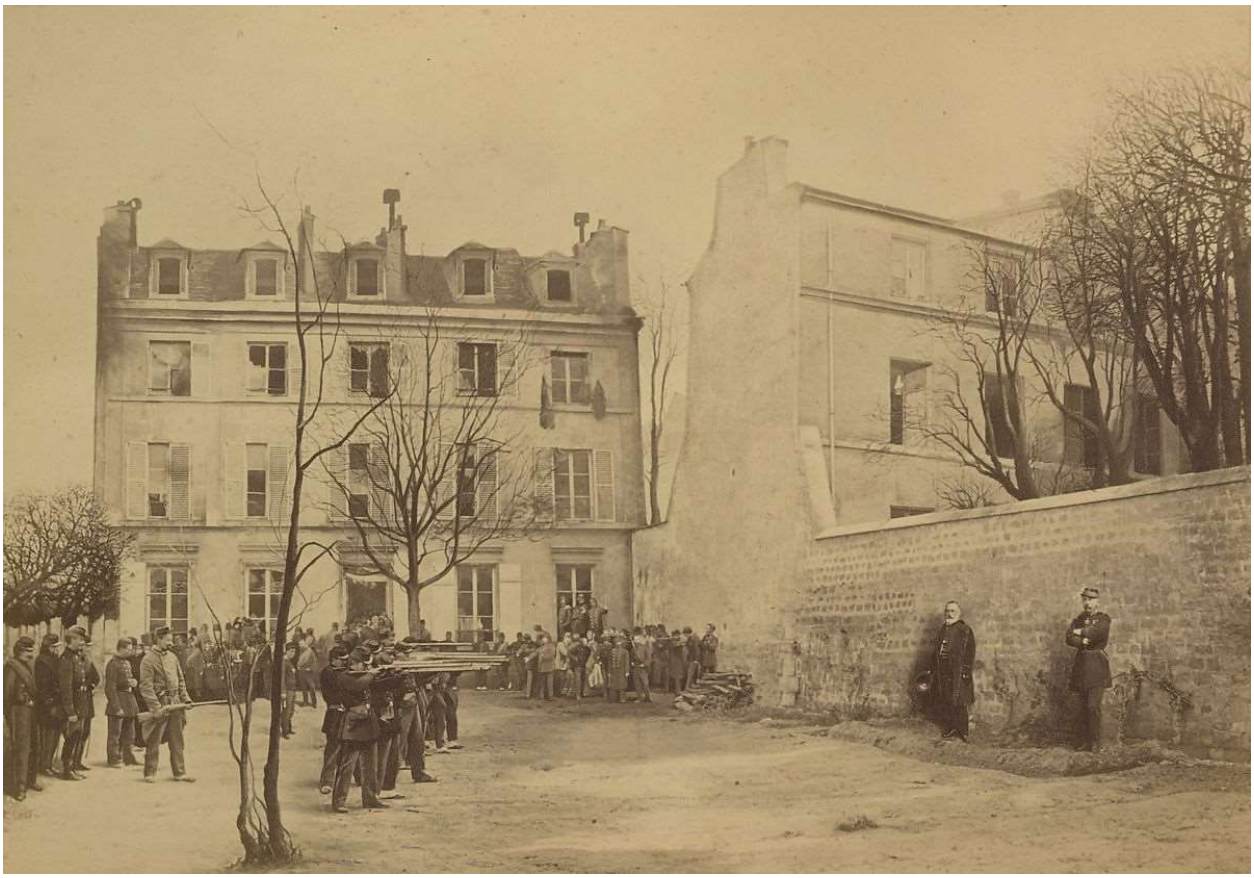

Figure 3: Staged photographs of a firing squad with the victims pasted into the scene.[58]

One might expect that if the photographs were accepted as authentic, the rebel sentiment would die down. After all, these images portrayed the participants fairly negatively. However, the disinformation campaign did not achieve the desired results, and in fact, the French government actually requested that the images be taken out of circulation because they were stirring anti-government sentiment.[59] This may have been precisely because viewers understood that the images were politically biased and inflammatory.[60] In other words, regardless of whether the viewer understood the image to be an authentic record of events, supporters of the rebels knew that the pictures served as propaganda and did not illustrate, at least fully, the reality on the ground.

However, as institutions began giving credibility to these images, a shift began to occur in our collective understanding of the medium. For example, the work of Mathew Brady during the American Civil War allowed viewers to see some representation of its realities. Already an established photographer, Brady captured the devastating loss of the union army at the Battle of Bull Run.[61] With a visual account that was able to mimic the battle as it truly happened, the depictions of war correspondents became less necessary.[62] One photographic magazine, *Humphrey's Journal*, went as far as to say that the photos

[58] Ernest Eugène Appert, "Souvenir 1870-71," The Metropolitan Museum of Art, last visited July 24, 2019. https://bit.ly/2JYAFDD.
[59] Press Event, "Faking It: Manipulated Photography before Photoshop," National Gallery of Art, July 1, 2013. https://bit.ly/2YmIjRp.
[60] Ibid.
[61] AMERICAN PHOTOJOURNALISM, p. 50.
[62] Ibid.

served as the only reliable records at Bull Run: "The correspondents of the Rebel newspapers are sheer falsifiers; the correspondents of the Northern journals are not to be depended upon, and the correspondents of the English press are altogether worse than either; but Brady never misrepresents."[63] What is striking about such a sentiment is its resonance today, as essentially it suggests that the reporting of the media is biased and thus the photograph at least allows the viewer to judge the information more objectively and without slant. Unfortunately, these too were staged.[64] But the narrative surrounding Brady's role as a reporter rather than an interpreter indicates why society may have begun to shift their perception of photography.

And, as technology improved, the use of photographs as a source of news began to increase. The most notable development was the introduction of the 35mm Leica camera, as well as the development of technology that allowed these images to be printed on the page quickly.[65] No longer did photographers need to haul cumbersome equipment to a scene to get a shot, and as a result, they could be there more immediately as the events unfolded.

With this increased access by photographers, the use of photographs shifted from a representation of a given story to the story itself, mainly in the form of picture magazines.[66] A part of this growing trend was a push to depict life as it occurred.[67] Notably, Stefan Lorent, a Hungarian editor of two major German magazines drove this ideology, explaining:

> [T]he photograph should not be posed; that the camera should be treated like a notebook of the trained reporter, which records contemporary events as they happen without trying to stop them to make a picture; that people should be photographed as they really are and not as they would like to appear; that photo-reportage should concern itself with men and women of every kind and not simply with a small social clique; that everyday life should be portrayed in a realistic unself-conscious way.[68]

This became pervasive in American culture, with magazines like *Life* and *Time* releasing photographs of events as they happened. Most notably, this ideology of authentically recording events had the most impact during times of war. Legendary photojournalist W. Gene Smith recorded the lives of soldiers under fire during the battle in Okinawa.[69] During his work, he captured an exploding mortar shell an instant before its shrapnel injured him, and in fact, it was later recounted of him that: "Before he passed out, he requested a pencil and paper and wrote instructions about where to send his film."[70]

Interestingly, then, it was never the inherent nature of photography, at least in isolation, that drove trust in the medium. Rather, it was the development of societal and institutional preferences or expectations around photography and its ability to capture life without comment.

Of course, this doesn't mean that deception and manipulation went away. As society began to trust the images more implicitly, photography created more opportunities for manipulation to occur. For example, a photograph depicted a Canadian soldier thumbing his nose at Germans from atop a trench as his fellow soldiers attacked.[71] The rifles, however, still had their breach covers on, indicating that this

---

[63] Ibid, p. 51
[64] Ibid.
[65] *See generally* Ibid.
[66] Ibid, p. 141.
[67] Ibid, p. 142.
[68] Ibid; from an uncited in *Modern Photography* as quoted in John R. Whiting, PHOTOGRAPHY IS A LANGUAGE (New York: Zipp Davis Publishing, 1946), p. 22.
[69] Ibid.
[70] Ibid.
[71] Ibid, Fig. 9.

was simply a training exercise.[72] At this point, however, it was not the photos themselves that people appeared to be interested in, but rather the story they suggested. In fact, generals often invited photographers to such events, and then photographers developed intense stories about the dangers they went through to get the shots.[73] These images accurately depicted the scene itself, but was simply made up by the photographers.[74]

And many photographers and editors clearly manipulated photographs as a means of conveying a message that supports a particular viewpoint, much like image based memes operate today.[75] These types of images didn't adhere to the truth, but played on existing beliefs and notions to connect with a target audience, making no attempt to depict events as they occurred.[76]

However, regardless of intent, it is clear that even as society shifted to an understanding of photographs and images as authentic, manipulation still occurred; the camera did lie. The specific technology itself was never a source of truth of its own accord. Rather, it was the institutional adherence to the principle of integrity and a general belief that photographs should depict events as they occur that drove trust in the medium.

But what does this mean for deep fakes then? The premise is false. It is true that people generally believe what they see. And, for videos, this problem is even more profound as audio elements are introduced, as well as the perceived difficulty in falsifying video rather than still images. However, rather than solely a destruction in our inherent human understanding, it is the destruction of a norm that has developed over time. This means that although fake video will indeed cause harms, society will be able to adapt to them as societal norms evolve.

A significant part of this is the existence of similar institutional adherence to truth, much like Time and Life. These societal norms can and will continue to drive trust in video as the viewer will understand that these institutions investigated the claims beyond just what appears on screen. And to the extent that videos become more consistently faked, society will shift back towards looking at the context behind the video. Who shared it? What was the original source? Does this seem plausible? Obviously, this doesn't mean that no harms will occur, and as discussed below, the internet presents a unique challenge that early photographic manipulation didn't, but there is reason to believe that this shift in approach will limit, at least to some extent, the harms that deep fake media will cause.

b. The Rise of Digital Manipulations and Industry Responses

These projections about deep fakes, which will be discussed in more detail in Section IV, are not mere speculation. Society has already dealt with similar developments altering this "seeing is believing" adage in the past. Deep fakes will likely progress similarly to these developments.

When digital photographic manipulation techniques began to commercialize in the 1990's, it was no surprise to hear the worries from some who now feared that the medium could no longer be trusted. But these fears never truly came to pass. Sure, individuals are still tricked into believing that a fake photo is authentic, but the widespread fear that we could no longer believe what we see didn't materialize for a few reasons.

For one thing, society caught on to the technology. This is partly because many early digital photographic manipulations were obviously bad and easily detected, to the point that they became

[72] Ibid.
[73] Ibid, p. 89.
[74] Ibid.
[75] FAKING IT, p. 100.
[76] Ibid.

popular targets of derision.[77] It is also likely because the prevalence of photo manipulation of fashion models sparked a national conversation about body image.[78] While similar mistakes could and did occur before digital editing, the accessibility of the tools and the lack of skill required to use them meant that amateurs could publish the content.

More importantly, the photojournalism community has taken steps to increase the trust that consumers have in their platform. The Associated Press, for example, has established policies on the limits of altering photographs. "The content of a photograph must not be altered in Photoshop or by any other means. No element should be digitally added to or subtracted from any photograph."[79] Because the focus remains on reinforcing consumer trust in the outlet, minor adjustments such as cropping, dodging and burning, conversion into grayscale, and normal toning and color adjustments (limited to those minimally necessary for clear and accurate reproduction) do not raise the same concerns as those that alter the content and therefore are allowed by the Associated Press.[80] These guidelines set a baseline for viewers of the photographs, allowing them to understand the media and how AP altered it. Because the photographs are tied to the journalistic integrity of the newspaper, AP self-regulated to ensure that consumers would still trust in their reporting.

Likewise, photographers have a strong incentive to ensure that the photographs that they capture and send to newspapers depicts the subject as it occurs. If the photographer cannot be trusted, his or her photos will not be printed. As a result, industry associations have established their own code of ethics as well.

The National Press Photographers Association, which is a professional society that promotes the highest standards in visual journalism, produced their own code of ethics designed to "promote the highest quality in all forms of visual journalism and to strengthen public confidence in the profession."[81] Such a code of conduct is vital because if their photos cannot be trusted by a newspaper, the newspaper will simply look to other photojournalists with a better reputation.

For example, an American war reporter deceptively edited a photograph from the 2003 invasion of Iraq to make it appear as though a British soldier was directing a man holding a child to take cover from incoming fire.[82] The photo didn't drastically change the outcome of the image, but ultimately made it appear as though a soldier's hand gesture was directed to an individual carrying a young child.

---

[77] For example, in 1994, Time Magazine ran a mugshot of O.J. Simpson, but darkened the image. Because this ran side-by-side with other magazines who did not darken the photo, the changes were apparent to the audience. "Altered Images," *Bronx Documentary Center*, last visited July 24, 2019. http://www.alteredimagesbdc.org/oj-simpson/.

[78] Carrie Arnold, "What's Photoshop got to do with it?" *Psychology Today*, June 29, 2011. https://bit.ly/2XPjbmU.

[79] "Code of Ethics for Photojournalists," *Associated Press*, 2018. https://bit.ly/2YUugj1.

[80] Ibid.

[81] "Code of Ethics," *National Press Photographers Association*, 2018. https://goo.gl/gEZXwJ.

[82] Frank Van Riper, "Manipulating Truth, Losing Credibility," *Camera Works*, last visited July 24, 2019. https://wapo.st/1lYTy5b.

Figure 4: The first two images used to create the final, altered photo[83].

While the final image may not have diverted too far from the original image sources, it did in fact betray the trust of the viewer who expected the image to depict the situation as the photographer saw it. Therefore, when alerted to the fake, the Los Angeles Times decided to fire the photojournalist.[84] The need to maintain reader trust left them with little other choice. This type of strict management left the

[83] Ibid.
[84] Ibid.

Los Angeles Times to be named as a runner up on Forbes' 10 most trustworthy journalism brands and receive a High Factual Reporting rating from MediaBias/FactCheck.[85]

Do societal adaptation and media norms mean that no one falls for photo manipulation? Of course not. Many major news organizations including the L.A. Times, the New York Times, and a variety of others ran with a photo received from Agence France-Presse which depicted Iranian missile tests, with four rockets launching into the sky.[86] However, the photo originally came from the Iranian Revolutionary Guard and had digitally included the fourth rocket. As one editor explained, "As the media editor working the msnbc.com home page yesterday, I was frustrated with the quality of a fuzzy video image we published of the Iranian missile launch, so I was thrilled when the top image crossed the news wires."[87]

This type of mistake will happen, but the key is the response. Most organizations took steps to disown the photos the same day, including Agence France-Presse who published an article explaining the mistake and citing experts to explain the change.[88] It also raised doubts to the credibility of photos received from the Iranian government. As institutional actors engage in such campaigns, news organizations will be less likely to use such content in the first place. That is all to say that while society undoubtedly will see and believe fake content, institutional backstops will work in the background to limit these occurrences and the harms they create.

III. KEY DISTINCTIONS TO CONSIDER

The above section explains that media itself has never been solely a source of truth. Instead, individuals examine the surrounding context of content such as the message it sends and the institution that published it. However, as deep fakes develop, it is important to also consider differences between the information ecosystem of 2019 and the information of 1990 or 1850. Deep fakes, as a unique form of disinformation, cannot simply be viewed in isolation or compared to techniques of the past without the context of the information environment at the time. Rather, we must fully consider how content disseminates as well as the ease of which to produce the fakes may alter the societal response to the fake content.

a. Social Media

Unlike the examples of photographic manipulation above, deep fakes do not necessarily disseminate through the same avenues as photographs, namely magazines, newspapers, and television broadcasts. Instead, individuals can share the content with those in a social network without the institutional checks that remain for traditional content publication and dissemination.

And clearly these concerns have some merit. Disinformation and false claims spread more rapidly today because of the virality of the content, existing biases, filter bubbles, and general polarization of political ideology.[89] For example, with traditional photographic fakery, the images had to be shared through the institutions themselves. This mean that the institutions could more actively engage in investigative

[85] Paul Glader, "10 Journalism Brands Where You Find Real Facts Rather Than Alternative Facts," *Forbes*, Feb. 1, 2017. https://goo.gl/rXq8pq.
[86] Mike Nizza & Patrick J. Lyons, "In an Iranian Image, a Missile Too Many," *New York Times*, July 10, 2008.
[87] Ibid.
[88] Ibid.
[89] Robert Chesney & Danielle Citron, "Deep Fakes: A Looming Crisis for National Security. Democracy and Privacy?" *Lawfare*, Feb. 21, 2018. https://bit.ly/2EP4nvf.

analysis and editorial discretion. With the ability for anyone to share forged images, these institutions may have less ability to limit disinformation.

Deep fakes, as a form of disinformation will definitely be able to exploit these phenomena to have an impact not entirely comparable to photography and historical trends. However, the concerns surrounding deep fakes are not simply that the new medium can exploit these phenomena, but rather that the apparent authenticity of the medium will drive belief and traditional institutions will fail to provide checks on this new medium. To the extent that people fear social media will allow deep fake content to spread independently of the institutions, mainly due to their apparent authenticity sparking trust and sharing, there are two reasons to think the impact may not rise to the level that some fear.

First, research indicates that much of the "filter bubbles" that people worry about with social media stem not from the technology, but rather the institutions themselves.[90] With the rise of social media, many argue that allowing users to artificially select the content they see drives them into an ideological rabbit hole.[91] These individuals interact only with those who are likeminded, further driving them to the political extremes and foreclosing on the ability to accept differing ideas.[92] However, research indicates that technology nor the internet does not drive these filter bubbles, but rather institutional media outlets.[93] As the researchers explain, "… the introduction of the internet and social media does not itself put pressure on democracy as such."[94] While challenging, this indicates that the internet may not have as significant an effect on the spread of deep fakes as some fear, but only so long as society desires accurate information rather than information that conforms to existing beliefs.[95]

Second, the apparent authenticity of fake media content matters less than the psychological factors that drive trust and sharing online. The human mind doesn't need much convincing to believe disinformation,[96] especially when the information comes from those we trust.[97] This is compounded when the individual actually wants to believe the information presented to them. If a deep video depicts information that contradicts an individual's pre-existing beliefs, the individual will either contort the facts to align with the existing beliefs or reject the facts outright. For example, President Trump's press secretary made the claim the more people attended his inauguration than any other inauguration in

---

[90] Yochai Benkler, Robert Faris, and Hal Roberts, NETWORK PROPAGANDA: MANIPULATION, DISINFORMATION, AND RADICALIZATION IN AMERICAN POLITICS (Oxford Scholarship Online 2018), p. 383. https://www.oxfordscholarship.com/view/10.1093/oso/9780190923624.001.0001/oso-9780190923624.

[91] Seth Flaxman, Sharad Geol, and Justin M. Rao, "Filter Bubbles, Echo Chambers, and Online News Consumption," Public Opinion Quarterly 80, Mar. 22, 2016, pp. 298-300. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2363701.

[92] Ibid.

[93] Benkler, supra n. 94.

[94] Ibid, p. 386.

[95] Ibid.

[96] Studies have shown that rudimentary disinformation can generate inaccurate memories. For example, researchers were able to implant fake childhood memories in subjects by providing a textual description of an event that never happened, making the person believe that they lived this false memory. While video could target more senses than an image or some other less engaging content, a simple textual description of events alone sufficed to implant the fake memories in the subject's mind. Elizabeth Loftus, "Creating False Memories," *Scientific American* 227, 1997, pp. 70-75. https://bit.ly/2GfjtbR.

[97] In a study conducted by the America Press Institute, researchers created a simulated Facebook post about recent health news and presented it to an online sample of 1,489 U.S. adults. The researchers presented the post as being shared from one of eight public figures and written by either the Associated Press or a fictional source. When asked about the relative pieces, participants were more likely to trust the source when the article had been shared by more trustworthy sources. "'Who shared it?' How American's decide what news to trust on social media," *American Press Institute*, Mar. 20, 2017.

history.[98] Despite the video evidence and a side by side comparison indicating the contrary, many supporters identified a photo with a lower turnout to show a fuller crowd because they knew it was a photo of the Presidents inauguration.[99] In other words, the individuals either convinced themselves that the crowd size was larger despite observable evidence to the contrary, or knowingly lied to support a partisan position. Unfortunately, this also means that malicious actors do not need high quality deep fake media to deceive a target audience. Instead, similar results can be achieved by more rudimentary methods because consumers existing beliefs will drive whether they believe in the information presented.

Even if a video isn't believed by the target audience, there is still damage that could be done by the wide-spread propagation of the content. Simply viewing the content can reinforce beliefs that the user already had even though the individual knows the content is an exaggeration or a parody. Further, realistic video may also be more tempting to share for some users. As the popularity increases, so too does the credibility as it reaches the edges of a network.

Nevertheless, there are several reasons to believe that deep fakes will not be significantly more problematic than any other form of media shared online. First, news stories can target the specific emotional factors that drive content sharing online even without the apparent richness of a deep fake video.[100] Image based memes, likewise, already spread at alarming rates due to their simplicity and ease of conveying information.[101] Therefore, a realistic fake video which can target the specific emotions of the user and the user's existing beliefs will spread rapidly online, but so too does a news article or a simple image based meme. Second, the herd behavior tied to sharing and popularity of content applies to any form of media, not just a realistic deep fakes.[102] Therefore, if users begin sharing a written story or any other "fake news" content, more users will in turn share that content regardless of the medium.

Indeed, deep fakes present new challenges and can exploit modern networks to transverse the globe at rapid rates. And this section isn't meant to imply that social media and the Internet do not pose unique challenges to modern disinformation. However, the challenge they present does not differ significantly than any other form of disinformation, and while the new technology presents a unique challenge, we shouldn't assume that it will drive the doomsday scenarios some fear: Society will likely adapt to these phenomena as they have to other new technologies in the past.

b. Accessibility of the Technology

Unlike the early days of photography, anyone can create a deep fake video without the need for advanced understanding of the techniques for forging images.[103] The core AI system is open source,[104] and many user facing applications allow anyone to input a small amount of training data and create deep

---

[98] Brian F. Schaffner & Samantha Luks, "This is what Trump voters said when asked to compare his inauguration crowd with Obama's," *Washington Post*, Jan. 25, 2017. https://www.washingtonpost.com/news/monkey-cage/wp/2017/01/25/we-asked-people-which-inauguration-crowd-was-bigger-heres-what-they-said/?utm_term=.0e49877243cf.
[99] Ibid.
[100] Vian Baker & Andrew McStay, "Fake News and the Economy of Emotions," *Digital Journalism* 6:2, 2018, p. 14. https://bit.ly/2JwXijB.
[101] Arturo Deza & Devi Parikh, "Understanding Image Virality," *2015 IEEE Conference on Computer Vision and Pattern Recognition*, June 2015, p. 1818. https://bit.ly/2LkFXMU.
[102] *See* Adrien Guille et. al., "Information Diffusion in Online Social Networks: A Survey," SIGMOD Record 42:2, 2013. https://bit.ly/2wFYs4X.
[103] John Villasenor, "Artificial Intelligence, deepfakes, and the uncertain future of truth," *Brookings Techtank*, Feb. 14, 2019. https://brook.gs/2TLjGbK.
[104] TensorFlow. https://www.tensorflow.org/.

fake content.[105] Indeed, more deep fake videos will be developed and shared than quality forged images. However there are still a few reasons why this may not present the significant challenge that some fear, at least in the context of disinformation.

First, these applications are still far from perfect, and many of the user-generated fake videos leave glaring mistakes.[106] As with photoshop before it, as the technology gets into the hands of amateurs, the nature of the tech becomes clearer to all those who see it because of lower quality that become the target of mockery or fail to achieve a political goal.

Second, having amateurs using the technology also demystifies it. Many still view deep fakes as almost magic, with no context for how the content is created. However, as individuals can create their own fake images, the technology becomes just another tool. And as the nature of the technology is understood, the serious impacts it could cause become more limited.

IV. SOCIETAL REACTIONS TO DEEP FAKES

Given the potential harms that can arise through the spread of deep fake generated audio and video, the question becomes how society will adapt to the technology and whether a regulatory intervention is necessary.

In view of the above critique of concerns about the impact of deep fakes and in view of the historical responses to digital photo manipulation, the answer appears to be that market-based solutions will likely be sufficient. As specific regulations overseeing deep fake technology, specifically calls to amend Section 230, could limit the significant benefits that the technology can provide, policymakers should proceed cautiously and narrowly tailor any action to a specific harm.

a. Expected Societal Responses

There are multiple ways we can expect society, and the market generally, to adapt to deep fake content.

Individual Adaptation to the Technology

In Section III, this paper explains that photographs have always been fabricated. Despite this, people have come to expect, at least to some degree, that photographs will serve as authentic record of events, especially when supported by institutions that support veracity. The increased access to digital editing tools meant more individuals could edit and share photos. Despite this phenomenon, society adapted to the changes.

Deep fakes will likely proceed on a similar path, with faked videos achieving disinformation goals, but over time the impact will gradually decrease as more people become aware that videos can and will be manipulated. The first step in this process may have been the public service announcement of President Barack Obama, which was actually a deep fake video with comedian Jordan Peele mimicking the voice of the former President.[107] Further, many fake videos spur controversy, and this controversy spurs awareness in turn. The Nancy Pelosi video, a so called cheap fake, achieved little harm due to quick

[105] *See, e.g.* "DeepfakesApp" https://deepfakesapp.online/.
[106] Jeffrey Westling, "Deep Fakes: Let's Not Go Off The Deep End," *Techdirt*, Jan. 30, 2019. https://www.techdirt.com/articles/20190128/13215341478/deep-fakes-lets-not-go-off-deep-end.shtml.
[107] David Mack, "This PSA About Fake News From Barack Obama Is Not What It Appears," *Buzzfeed News*, Apr. 17, 2018. https://bit.ly/2E31FCg.

debunking, but led to national coverage about video editing and the potential for deep fake videos to impact elections.[108]

At the same time, just because people understand that video can be manipulated doesn't mean that they will distrust all videos. Much like with the rise of photojournalism, content will still be judged, at least to some extent, by the context surrounding it.[109]

And while the psychological factors described in Section III and social media virality may lessen the strength of the institutional checks on disinformation using inauthentic video, it is a problem for disinformation as a whole separate and distinct from deep fakes. To the extent that people want to consume information that conforms to their pre-existing world views, a deep fake video is unnecessary to achieve those harms.

Institutional Actors and Independent Debunking of Deep Fake Content

It is true that the internet has allowed for the development of more such actors, and this influx of voices may lead to increased disinformation spreading online, but the doomsday scenario that some predict will likely never occur because many institutional actors still rely heavily on the trust of their audience, and as such will investigate heavily the claims any video may appear to show. This includes both forensic analysis of individual content as well as traditional investigatory work such as examining sources and interviewing subjects/witnesses of the event.

Further, there is a strong market incentive for news organizations to debunk fake videos. When the recent video of Nancy Pelosi was posted online, the debunking of the video received much more attention than the original video ever received.[110]

In fact, websites like Politifact and Snopes have already developed an entire business model based upon checking the veracity of statements made by politicians and the news.[111] In a similar vein, independent "authenticity checkers" could identify and bring to light fake videos that are published by news agencies or shared over social media platforms. For example, many organizations debunked a compressed version of CNN's Jim Acosta "assaulting" a Whitehouse intern, showing how the poor quality of the White House's version of the video made it appear as though the reporter was more aggressive than he actually was.[112] Likewise, a slowed down video of Nancy Pelosi was quickly debunked and scrutinized, leading towards more public awareness about video manipulation.[113] This type of investigative reporting can significantly limit the impact of harmful deep fake content, while also creating an entire competitive industry.

Perhaps more importantly, using inauthentic media presents a unique opportunity for these institutions dedicated to conveying accurate information or debunking disinformation. To the extent that they claim a given situation occurred without any witnesses to that fact, most fake news stories are not inherently

---

[108] Jeffrey Westling, "Fool me once… You can't get fooled again," *Morning Consult*, June 3, 2019. https://bit.ly/2WsP9EK.
[109] Jeffrey Westling, "Deep Fakes: Let's Not Go Off The Deep End," *Techdirt*, Jan. 30, 2019. https://www.techdirt.com/articles/20190128/13215341478/deep-fakes-lets-not-go-off-deep-end.shtml.
[110] *Ibid.*
[111] "Politifact," *Politifact*, last visited July 24, 2019. https://www.politifact.com/; Fauxtography, *Snopes*, last visited July 24, 2019. https://www.snopes.com/fact-check/category/photos/.
[112] *See* Aymann Ismail, "The White House's Acosta video looks different from the original. Does that mean it's 'doctored'?" *Slate*, Nov. 08, 2018; *see also* Charlie Warzel, "Welcome to the dystopia: People are arguing about whether this Trump press conference video is doctored," *Buzzfeed News*, Nov. 8, 2018. https://bit.ly/2D9UAO5.
[113] Drew Harwell, "Faked Pelosi videos, slowed to make her appear drunk, spread across social media," *Washington Post*, May 24, 2019. https://wapo.st/2LWzOWC.

falsifiable. An investigative report can explain how the accusation has no support, those who wish to believe the story can explain this away, simply stating that the "perpetrators" covered up the evidence.

Fake video can currently be debunked using technical means.[114] This makes them inherently falsifiable. With a narrow focus on debunking fake videos, bad actors may simply choose to produce these false stories rather than risking outright debunking. And to the extent that the video begins to spread online, platforms would be more likely to remove blatantly manipulated images and video rather than a story citing unnamed sources that cannot be disproven.[115]

Existing Platform Response

Social media platforms and other avenues for the dissemination of deep fake content face at least some pressures, whether from the user or the advertisers, to root out fake content. On July 26, 2018, Facebook lost $119 billion of its value due to decline in growth of the active user-base.[116] While many factors played a role in this decline, the reports regarding Russian election interference on the platform could have played a major role.[117] In response to many of these claims, Facebook began taking steps such as shutting down its Trending News section and banning pages aimed at US election interference.[118] As with the more general issue of fake news, platforms will likely feel some pressure to take steps to limit the spread of potentially harmful deep fake videos as users may choose to utilize different platforms to connect with their social networks, and platforms have begun to invest in deep fake identification.

In fact, some social media platforms have already chosen to invest in and employ deep fake detection technology that can identify these videos and target them for review. Facebook, for example, has said that the company has been collaborating with academics to protect its users from real world harms caused by deep fake videos.[119] And unlike a more heavy-handed approach, if companies begin over-removing user content, the user-base may begin to move onto other platforms. This will in-turn drive a balance between removing harmful content and allowing free speech.

Interestingly, Reddit, a website composed of over a million of user generated communities and the birthplace of the term "deepfake," faced similar problems with the spread of deep fake pornographic videos. Because of the platforms structure, individual subreddits, created by users, became a place to consolidate and share these images and videos. In 2018, Reddit took steps to resolve these issues by updating the websites policy to prohibit "the dissemination of images or video depicting any person in a state of nudity or engaged in any act of sexual conduct apparently created or posted without their

---

[114] SUMMARY OF DISCUSSIONS AND NEXT STEP RECOMMENDATIONS FROM "MAL-USES OF AI-GENERATED SYNTHETIC MEDIA AND DEEPFAKES: *PRAGMATIC SOLUTIONS DISCOVERY CONVENING*," *Witness*, June 11, 2018. https://blog.witness.org/2018/07/deepfakes/.

[115] The most recent test case is the video of Nancy Pelosi. The video was slowed down to make her appear drunk, and Youtube and Twitter removed the video despite the relatively minor editing. Facebook choose to limit its visibility, but didn't remove the video outright. Barbara Ortutay, "Facebook isn't deleting the fake Pelosi video. Should it?" *Associated Press*, May 30, 2019. https://www.apnews.com/83f2e0438b8d45249dcc05f4fb430f64.

[116] Tom Metcalf, "Zuckerberg Loses $16.8 Billion in a Snap as Facebook Plunges," Bloomberg, July 25, 2018. https://goo.gl/L6Csuf.

[117] David Chau, "Facebook share price drop wipes $US119b from company's value, $15b from Mark Zuckerberg's net worth," ABC, July 27, 2018. http://www.abc.net.au/news/2018-07-27/facebook-share-price-drop-wipes-$us119-billion-company-value/10042404.

[118] Aja Romano, Mark Zuckerberg Lays out Facebook's 3-Pronged Approach to Fake News, Vox, April 3, 2018. https://www.vox.com/technology/2018/4/3/17188332/zuckerberg-kinds-of-fake-news-facebook-making-progress.

[119] Sara Ashley O'Brien, "Deepfakes are coming. Is Big Tech Ready?," CNN Money, Aug. 10, 2018. https://goo.gl/p8LaRx.

permission, including depictions that have been faked."[120] This eventually led to the banning of communities such as /r/deepfakes, and admins removing such content when reported.

That is not to say that no harmful deep fake content will be shared online, and many sites will still have some incentive to host such content.[121] And there is an incentive for platforms to encourage inflammatory with ad-based businesses. However, for the most damaging content, specifically those that target civic discourse and society as a whole, market pressures from both users and advertisers will drive platforms to limit their spread and dissemination, especially those platforms with a large, diverse user base.

Ongoing Deep Fake Detection Research

To respond to deep fakes, companies must develop and employ strategies to identify the specific videos. However, some have raised concerns about whether deep fake videos would actually be identifiable as the technology continues to improve.

Academia, working with The Defense Advanced Research Projects Agency, has already made significant strides on this front.[122] Researchers from the University of Albany use blinking to detect whether an AI generated facial videos.[123] The approach attempts to detect the "physiological signals intrinsic to human beings that are not well captured in the synthesized videos. Such signals may include spontaneous and involuntary physiological activities such as breathing, pulse, and eye movement…"[124] Similarly, MIT researchers demonstrated the ability to use AI to take someone's heart rate from video.[125] Because deep fakes do not have heartbeats, this type of analysis may be used to differentiate the fakes from authentic videos. And even more basic issues like altering the resolution of the generated video to fit with the target video leaves detectable differences.[126]

Ultimately, the AI technology will continue to improve and the forgeries will continue adapting to the detection software. But that has always been the case. Though the advancement of the technology used to fake videos will continue, so will the technology used to detect it. In fact, leading researcher in the field Siwei Lyu recently explained that beyond the blinking analysis his team has published, there are more techniques that remain unpublished, allowing detection to stay one step ahead of the forgeries.[127]

---

[120] /u/landoflobsters, "Update on site-wide rules regarding involuntary pornography and the sexualization of minors," *Reddit*, Feb. 7, 2018. https://www.reddit.com/r/announcements/comments/7vxzrb/update_on_sitewide_rules_regarding_involuntary/.

[121] Alice Marwick & Rebecca Lewis, "Media Manipulation and Disinformation Online," *Data & Society*, 2017, p. 42. https://bit.ly/2XKrHDO.

[122] Will Knight, "The US military is funding an effort to catch deepfakes and other AI trickery," *MIT Technology Review*, May 23, 2018. https://goo.gl/w9u6th.

[123] Yuezun Li et al., "In Ictu Oculi: Exposing AI Generate Fake Videos by Detecting Eye Blinking," June 2018. https://arxiv.org/pdf/1806.02877.pdf.

[124] *Ibid.*

[125] Will Knight, "The US military is funding an effort to catch deepfakes and other AI trickery," *MIT Technology Review*, May 23, 2018. https://goo.gl/w9u6th.

[126] Yuezin Li and Siwei Lyu, Exposing Deepfake Videos by Detecting Face Warping Artifacts, Nov. 1, 2018. https://arxiv.org/pdf/1811.00656.pdf.

[127] Will Knight, "The Defense Department has Produced the First Tools for Catching Deepfakes," *MIT Technology Review*, Aug. 7, 2018. https://goo.gl/rfqr85.

And, as noted above, private firms are already investing in technologies to identify and detect deep fake videos.[128]

As of 2018, every generation model had a powerful detection tool.[129] And many researchers believe detection will stay ahead of generation in the game of cat and mouse.[130] There is no doubt that generation techniques will improve, and other researchers worry that generation technology may develop past detection models, but it is important to understand this dynamic is constantly developing as new techniques for generation and detection develop.[131]

b. Regulatory Approaches

As society adapts to the technology, the market will address many harms deep fakes can cause to some extent, but this isn't a binary challenge. Rather, there is a spectrum of harms that can occur. To the extent that government action can minimize the harms caused by deep fake media, some regulatory responses may make sense. However, regulators should carefully consider the cost of such action, as well as the relative benefits, before making broad changes to the regulatory structure governing deep fake media.

Promoting Digital Literacy

An obvious but too little discussed alternative response to overregulation is simply education. A large part of the reason why Photoshop never became the 'death of trust' was because people broadly became aware of the technology, as well as the myriad ways that people may be manipulated by it.[132] And, in fact, with respect to deep fakes, many private entities have already made great strides on this front.[133] If regulators want to act in response to deep fakes, a good way to do so is to focus their effort on additional avenues to educate the public about what this technology is capable of. These efforts can help alleviate harms by making consumers aware that videos may not depict the truth they claim.

Increased Study

Depending on how a bill might be structured, governmental bodies can encourage and support research on the technology and its uses. A recent proposal, for example, would require the Department of Homeland Security to release a report on the state of the technology every eighteen months.[134] This type of report could help layout the current state of the field, providing lawmakers and industry with the necessary information to adequately target the harms deep fakes can cause. Likewise, DARPA and other Governmental entities have, and can continue to, support research into deep fake detection. These steps will impose relatively limited costs while providing significant benefits for both government and market responses to deep fakes.

---

[128] Aja Romano, Mark Zuckerberg Lays out Facebook's 3-Pronged Approach to Fake News, Vox, Apr. 3, 2018. https://www.vox.com/technology/2018/4/3/17188332/zuckerberg-kinds-of-fake-news-facebook-making-progress.

[129] SUMMARY OF DISCUSSIONS AND NEXT STEP RECOMMENDATIONS FROM "MAL-USES OF AI-GENERATED SYNTHETIC MEDIA AND DEEPFAKES: *PRAGMATIC SOLUTIONS DISCOVERY CONVENING*," *Witness*, June 11, 2018. https://blog.witness.org/2018/07/deepfakes/.

[130] Ibid.

[131] Ibid.

[132] Jeffrey Westling, "Deep Fakes: Let's Not Go Off The Deep End," *Techdirt*, Jan. 30, 2019. https://www.techdirt.com/articles/20190128/13215341478/deep-fakes-lets-not-go-off-deep-end.shtml.

[133] David Mack, "This PSA About Fake News From Barack Obama Is Not What It Appears," *Buzzfeed News*, Apr. 17, 2018. https://bit.ly/2E31FCg.

[134] Deepfake Report Act, H.R. 3600 (2019).

Technical Restrictions

Some have argued that the law should require the technology to include watermarks or other distinguishing characteristics so that investigators can determine more rapidly.[135] These considerations stem from the idea that the apparent authenticity of deep fakes will make detection difficult for even forensic experts. To some extent, this can provide benefits. News agencies can look to internal records showing whether an image or video has been manipulated. Or perhaps a more glaring indication that content is fake can help alleviate sharing on social media.

However, the costs here need to be weighed as well. Many legitimate uses for deep fakes can suffer, depending on how such technical restrictions are implemented. For example, one bill would essentially require movies using the technology to generate a historical figure to have clear disclosure during the movie, ruining any chance for the audience to suspend their disbelief.[136] And audio content would be required to have an interruption every two minutes.[137] Further, to the extent that news agencies couldn't just look at a the underlying "fingerprint," traditional methods of validation such as speaking to witnesses or investigating timelines still remain an option for most deep fake content.

Platform Liability

Under Section 230 of the Communications Decency Act ("CDA 230"), if a user posts an illegal deep fake video, the platform will not be liable to the third party as the publisher of the content. This policy makes sense. Without some immunity, platforms would not allow the users to have unfettered access to the forum.[138] Less ideas would be shared as controversial opinions may need to be filtered out.[139] Further, the potential costs would be felt proportionally more by smaller start-ups who try and compete with the incumbents who have the resources available to comply with over-bearing regulations.[140]

At the same time, modifying ("CDA 230") does potentially open up an avenue to stem the spread of deep fake videos. As explained above, the real problem with deep fakes, and disinformation generally, isn't the quality of the content, but rather the way we trust and share content online. As such, it is the ability for fake content to transverse a network at breakneck speeds that presents the serious concern. Restructuring Section 230 to target platforms could allow regulators to focus on the channels by which the videos spread.

However, there are serious questions as to whether a 230-based approach will in fact work to prevent the spread of harmful deep fake videos designed to cause societal harm, while imposing significant costs on free speech.

First, many of the most damaging deep fake videos aren't illegal except insofar as they constitute a tort on individual or entity—a private harm. In the case of a white officer appearing to shoot an unarmed black teen, only the specific officer or perhaps the police department could bring an action against the publisher of the defamatory content. Therefore, an amendment to Section 230 to treat a platform as a

---

[135] Sam Gregory, "Deepfakes and Synthetic Media: Updated Survey of Solutions against Malicious Usages," *Witness*, June 2019. https://blog.witness.org/2019/06/deepfakes-synthetic-media-updated-survey-solutions-malicious-usages/.
[136] Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019, H.R. 3230, § 1041(c).
[137] Ibid, § 1041(e).
[138] Cathy Gellis, "Section 230 Isn't About Facebook, Its About You," *Techdirt*, Feb. 26, 2018. https://goo.gl/FCCeU3.
[139] Ibid.
[140] "Sesta V. Fosta," *Engine*. https://goo.gl/9vMWqR.

publisher would do little in this case if the officer's face and badge are blurred out, despite the fact that such a video could be extraordinarily damaging to society—a public harm.

Second, any amendment to Section 230 would likely shift the focus away from moderating content that appears on platforms towards taking actions that simply shield the company from liability. For example, in the above example, because the potential deep fake isn't defamatory to any individual, platforms may not invest time and resources reviewing the content, instead focusing moderation on those videos that actually are illegal, even if they are not as harmful.

Third, legitimate speech will suffer. Without these protections, satire, parody, artistic and a wide variety of other deep fake content may be removed by moderators who fear that the content will leave platforms open to potential liability. Lawmakers must carefully consider such impact before pursuing any 230 based "solution" to deep fakes.

Criminalization

A final approach could be to impose criminal punishment on the generation of deep fake content. The potential costs here are significant, depending on how such an approach is structured. Criminal liability would drastically reduce borderline uses of the technology, such as the parody of a politician or public figure. For example, artists created a deep fake video of Mark Zuckerberg to highlight Facebook's decision not to remove an edited video of Speaker Nancy Pelosi.[141] Such a video could be considered defamatory, though unlikely under the New York Times v. Sullivan standard for public figures.[142] But because of the fear of criminal penalties, the individuals may in the future choose not to make such content.

However, if narrowly tailored and consistent with the first amendment, criminalization could drastically limit the harms of deep fakes. Lawmakers should very carefully consider whether such action would do more good than harm.

CONCLUSION

This paper has considered the nature of deep fake technology and its present and future applications. It has reviewed contemporary concerns about deep fakes being used in disinformation campaigns, and critiqued those concerns both in view of scientific research and based on historical precedents of technological deception. Finally, it has considered and discussed policy approaches to dealing with the problems that the technology may cause.

Regardless of the path, it is vital that any regulation is not based on the fear that deep fakes present a unique policy issue. As this paper has explained, deep fakes present similar challenges to other forms of deceptive media. However, the worst-case scenarios that some have outlined are now being used as a justification to pursue different policy changes. Some policy changes may be needed, but it is critical that the changes be based on actual harms and narrowly tailored to address these harms.

In a sense, some solace may be taken from this conclusion that deep fakes are in fact not so game-changing a technology as they may seem. For deliberative bodies such as governments, the possibility that malign technologies may outpace the ability to deal with them. Fear of that possibility can, and often has, resulted in impetuous efforts to legislate on or regulate those technologies without full consideration of the consequences. That even a technology as new and strange as deep fakes may in fact

[141] Samantha Cole, "This Deepfake of Mark Zuckerberg Tests Facebook's Fake Video Policies," *Motherboard*, June 11, 2019. https://www.vice.com/en_us/article/ywyxex/deepfake-of-mark-zuckerberg-facebook-fake-video-policy
[142] New York Times v. Sullivan, 376 US 254 (1964).

go the way of other technologies such as digital photo editing should help to quell these inclinations toward rashness, giving policymakers the time to hold to their moorings of first principles and consider the matter dispassionately with all deliberate speed.

ABOUT THE AUTHOR

Jeffrey Westling *is a Technology & Innovation Fellow with the R Street Institute. His research focuses on emerging technology and telecommunications. He has a B.S. in Ecology and Evolutionary Biology from the University of Arizona and received his J.D. from the University of Colorado Law School.*