# Deepfakes and Social Media

A Research Paper in STS 4600

Presented to the Faculty of the School of Engineering and Applied Sciences
Unviersity of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Engineering

Author
Siddharth Ghatti
May 08, 2020

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature Siddharth Ghatti Date 05/08/2020

Approved _____ Date _____

Rider Foley, Department of Engineering

## Introduction

A video begins circulating online of the American President discussing in private about how the Star Wars franchise is one of the worst movie franchises to ever exists. Star Wars fans everywhere are angered. They immediately take to social media platforms to share their discontent that the President of the United States does not share their passion for the film franchise. George Lucas himself shares a public statement sharing his disappointment that the president did not find the film franchise that he created to be noteworthy. It is later revealed that the video that angered an entire fan-base was nothing more than a proctored video using deepfake technology. The situation described is overall quite harmless. However, it becomes much more dangerous when the video that was proctored and released was instead one of the President formally declaring that the United States will use military force in North Korea. This video could cause increased tensions between both countries and their allies, damaged global markets and trade, and could have led to global hysteria. All of this is possible due to the fact that audio and video artifacts are often used as sources of indisputable evidence, a byproduct of the current difficulty in proctoring video and audio artifacts that are "good" enough to fool human beings. However, this norm is in jeopardy with the rise of deepfake technologies. Deepfakes are defined as fake video and audio artifacts that are generated by neural networks. Neural networks are computer algorithms that mirror how the brain works (Dack,2019). The products of deepfake technologies are often very realistic, to the point that Sean Dack (2019), an IPI Cybersecurity Fellow at the University of Washington, argues that deepfakes represent a revolution in how disinformation can be created. In the same article Dack also argues that the ease of manipulating social media makes it a key agent in disinformation campaigns, citing Iran's exploration of using social media to spread false information during the 2018 midterm

elections (Dack,2019). Using this information, it can be argued that a close relationship exists between deepfakes and social media. Social media provides an amiable platform to spread misinformation and deepfakes can be designed to be artifacts of misinformation. Together, these two tools can be used by malicious users to create havoc and forward political positions (Dack,2019). As a result, for my STS thesis I will analyze the potential impact that deepfake technologies will have on social media platform's content moderation methods. I aim to show that the threat that deepfake technologies present to information consumers will require a change in content management methods on social media platforms.

The realism of deepfakes is in large part due to how they are generated. Deepfakes are generated by Generative Adversarial Networks (GAN). A Generative Adversarial Network is comprised of a generator and a discriminator. The generator is designed to create images that are similar to real images. The discriminator, on the other hand, is trained to distinguish between real images and the artificial images. The generator works to minimize the probability that its images are detected as artificial by the discriminator while the discriminator works to maximize the probability that it can correctly identify images as authenticate or artificial. The GAN essentially has the discriminator guide the generator to create realistic images by having the discriminator give the generator information on what real images look like. As a result, by the end of this learning process the generator becomes adept at producing realistic images (Shen et al., 2018).

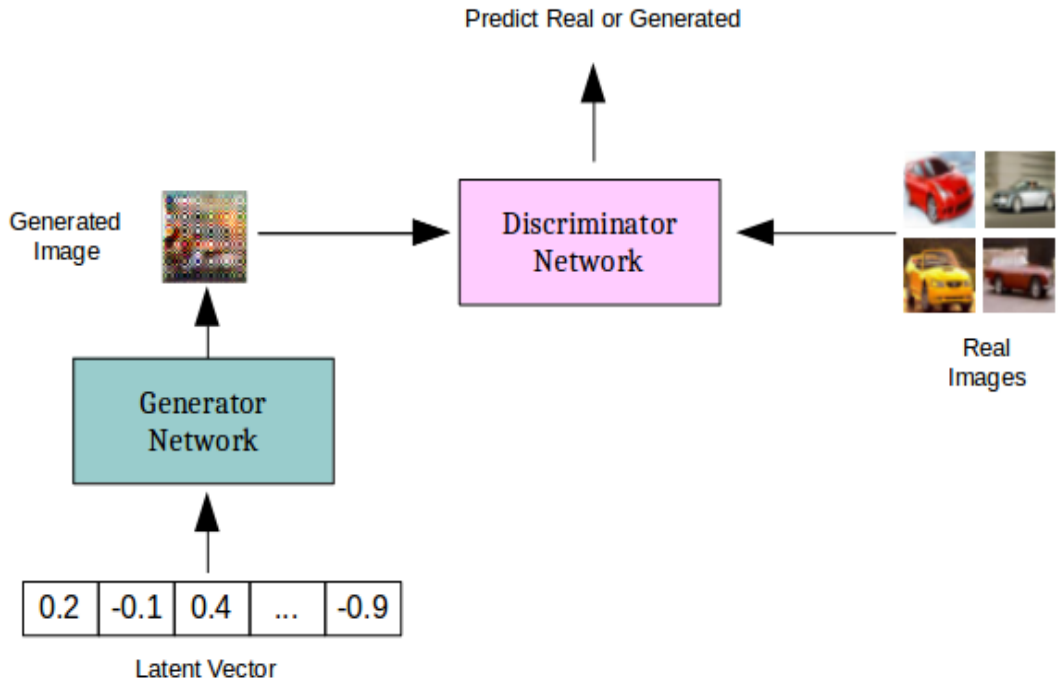


Figure 1. Generative Adversarial Network (Desai, 2018).

The realism of deepfakes created by GAN's creates an issue for social media platforms aiming to eliminate the spread of misinformation on their platforms. Currently, the content management work for social media platforms is done by moderators in countries such as the Philippines and India (Dwoskin, 2019). The current model of content moderation has humans make a judgment call on whether certain content can be allowed to remain on a platform. However, it can be seen that the current model fails with deepfake content. As this content, if it is "good" enough, can fool human beings into thinking it is real. This points to the need for social media platforms to deploy technological solutions to deal with deepfakes.

As a result, a case can be made for the delegation of certain content management to automated systems. Due to this, one STS framework that I will be using to analyze this issue will

be Actor-Network Theory (ANT). Specifically, I will be focusing on the delegation of labor to non-human agents (Latour, 1992).

**STS Framework**

Actor-Network Theory examines the changing relationships that exist between technology and society. Specifically, Actor-Network Theory builds upon the argument that sociotechnical systems are built through changing relationships between institutions and humans by positing that technological artifacts play an active role in these relationships as well. It argues that just as human actors have an effect on technology, technological artifacts impact human behavior as well. A more pertinent piece to my analysis of automated content moderation that comes from this theory is the concept of delegating labor to non-human agents. As the name suggests, this concept examines how certain tasks can be assigned to technology and whether the distribution of such responsibilities to technology is an appropriate action to take (Latour, 1992). It does this by looking at whether the technology is reliable, whether it discriminates across the user base, and whether it provides an appropriate program of action.

In addition to the technological complexities that social media platforms will have to address, there are also sociological complexities that social media platforms will have to wrestle with as well. If social media platforms took a stringent stand against fake content and over-actively banned content then issues over censorship would arise. This can be illustrated by the events that occurred in September of 2016. Journalist Tom Egeland included the famous picture of the "Napalm Girl", which illustrates children running away from a napalm attack, in an article that was about pictures that changed warfare. Egeland's Facebook post was deleted, and his Facebook account was suspended when he reposted the article twice. Given the image's iconic status, the actions taken by Facebook faced global criticism. After more than a week, the image

was reinstated (Gillespie, 2018). At the time many journalists accused Facebook of making the wrong decision. However, in hindsight, the issue was not that simple. Tarleton Gillespie (2018) argues that although the image is very important and should be seen by everyone, the content of the picture, nude children screaming in pain, is not one that is allowed in nearly all societies. The combination of the image's iconic status with its jarring content, content that would usually be banned on Facebook, created a dilemma for Facebook. Although the image of the "Napalm Girl" is not a deepfake, there would be similar implications and issues for any social media platforms when dealing with the censoring of deepfakes. On the other end of the spectrum, Facebook, more recently, has faced backlash for keeping a doctored video depicting House Speaker Nancy Pelosi seemingly slur her words on the platform. The reasoning behind this being that Facebook's content policy does not explicitly require the platform to ban content that is "False" (Scola,2019). The two cases above illustrate some of the social dilemmas that social media platforms will have to deal with when dealing with deepfakes. If these platforms take a stringent approach to deepfake content then criticism over censorship will rise. However, if they do not do anything to moderate deepfake content then a bleak future where information can be manufactured becomes a very real possibility.

**Research Question**

In hopes of avoiding this bleak future I hope to answer the question: How and why do social media platforms need to change their content management methodologies to prepare for deepfakes?

**Research Methods**

Since one of the major actors in the question are the current content management policies, one method of evidence collection that I will be using will be looking at policy

documents. Specifically, I will be looking at the current policies of the social media platforms Twitter, Facebook, and Reddit. In addition to this, in order to gauge the global communities current understanding of deepfake technology, I will utilize Google search data to estimate the number of internet users that know about deepfake technology. I will also continue to look at prior literature on the topic of deepfakes. I will be focusing on any research that is being done in the autonomous detection of deepfake videos. I will be using Actor Network Theory, specifically the delegation of labor to technology, to better analyze how some of these automated solutions can be effectively deployed. Finally, I will leverage research conducted on the spread of false news in order to understand how it can be applied to the spread of deepfakes.

**Results**

From the research that I have conducted I have found that there are several reasons why social media platforms need to change their current content management methods. The first is the fact that knowledge about what deepfakes are is still not widespread. The second reason is that false news spreads faster than true news (Vosoughi et al, 2018). The third reason is that Twitter, Reddit, and Facebook's current policies either explicitly or implicitly disallow deepfakes. In order to uphold these policies these social media platforms will have to rely on automated deepfake detectors.

*False News: Lessons from the Past*

In a study that was conducted in 2018 three researchers, Vosoughi, Roy, and Arai, conducted an analysis of false and "true" news instances that spread on Twitter to understand the differences between how each type of news instance spread. Using a dataset that included 126,000 instances of news that was spread by 3 million people more than 4.5 million times they found that false news spread much farther and faster than "true" news instances. On average it

took instances of “true” news six times longer to reach 1500 people than it did for instances of false news. Differences in how false news and “true” news spread was also found in research published in April 2020 by a different group led by Zilong Zhao. Using data from both Twitter and Weibo, a Chinese social networking platform, the publishing team found that the spread of false news instances and “true” news instances had different topological features starting in the very early stages. One such difference that the group found was that, initially, false news was not spread as much as true news. However, in later stages of spread would false news would outpace “true” news (Zhao et. al, 2020).

When it comes to user accounts that were spreading said news on the platform, Vosoughi, Roy, and Arai ‘s data showed that on average accounts that spread false news had a smaller number of followers and had been on the platform for less time. To understand the success that false news was having in spreading on the platform the team dissected the content of each news instance. They found that on average false news was much more novel than “true” news. Finally, by running multiple analyses with and without news instances spread by bots the team found that it was actually people and not bots that were more likely to spread false news (Vosoughi et al, 2018).

Some takeaways from this study that can be applied for deepfakes are:

1) The correlation between novelty of content and false news.
2) The characteristics of accounts that spread false news.
3) False news moves faster than true news.
4) The increased likelihood of humans to spread false news.
5) False news and “true” news spread differently.

*Deepfake Awareness*

Since its humans that are more likely to spread false news, I set out to understand how many internet users knew what deepfakes are. I did so by using the data available from Google Trends and looking at how many users were searching for deepfakes on Google. I decided to use Google for several reasons. First, I am operating under the assumption that social media users would use search engines to learn about deepfakes. Second, as of 2019, Google owns 92.7% of the market ("Search Engine Market Share in 2019", 2020). Third, Google is providing a public platform with their data through Google Trends.

Google Trends does not give absolute numbers for search words due to the increasing number of internet users over time. Rather, Google Trends provides a normalized view of the data by providing an indexed value ranging from 0-100 that represents that search word's interest in proportion to the maximum interest it generated on Google. 100 represents the maximum interest generated by the search word. This normalization allows for the comparison of values across different dates and regions (Rogers, 2016).

The first search term that I looked into using Google Trends was deepfakes. The results are illustrated below.

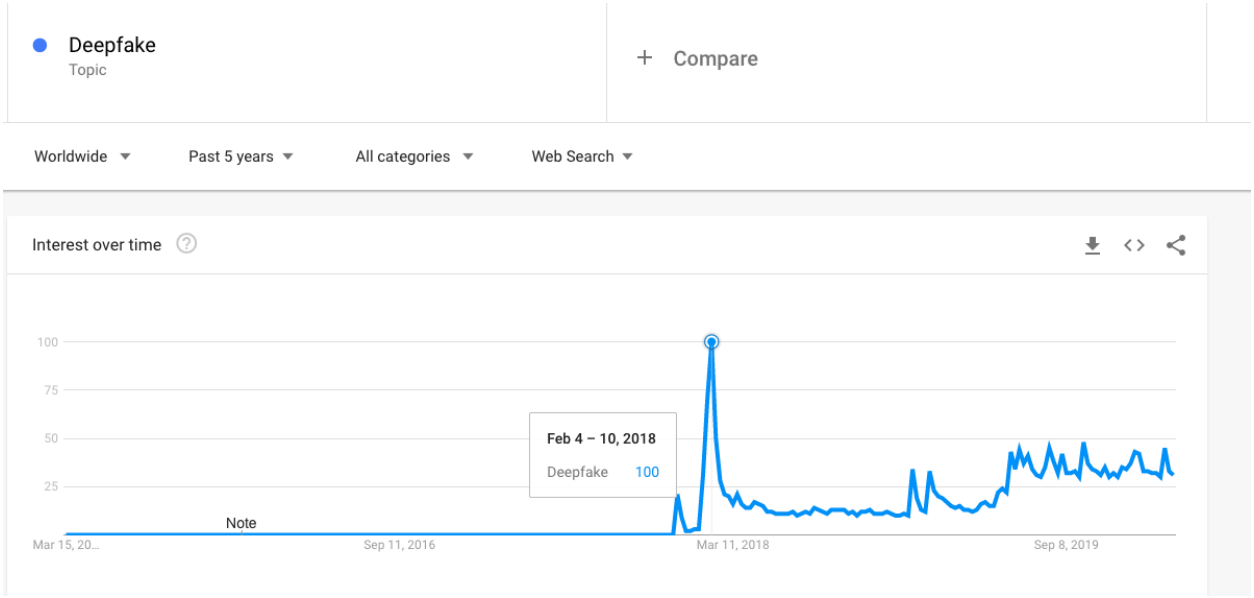

Figure 2. Deepfake Search Trend (Google Trends, 2020).

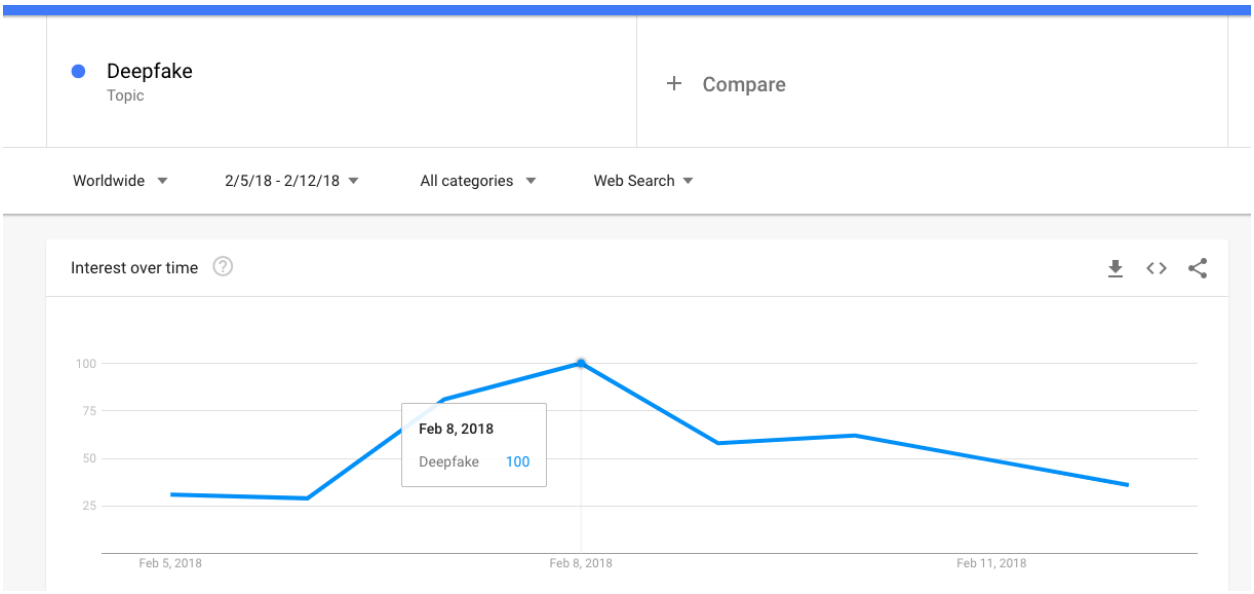


Figure 3. Deepfake Search Trend February 2018(Google Trends, 2020).

So, on February 8 2018 deepfakes was its most popular with a longer trend of popularity starting at the end of 2017 and rising once again in October 2019. The peak in February of 2018 can be attributed to reddit banning it's deepfake sub-reddit (Robertson, 2018), leading to an influx of media reporting on this move.

Overall this trend shows positive activity from users aiming to educate themselves about deepfakes. However, before we jump to positive conclusions, we must contextualize this data. One method that we can utilize is to compare deepfakes to phenomena that has global understanding, such as the (at the time of writing) on-going pandemic of Covid-19. The results of this comparison are below:

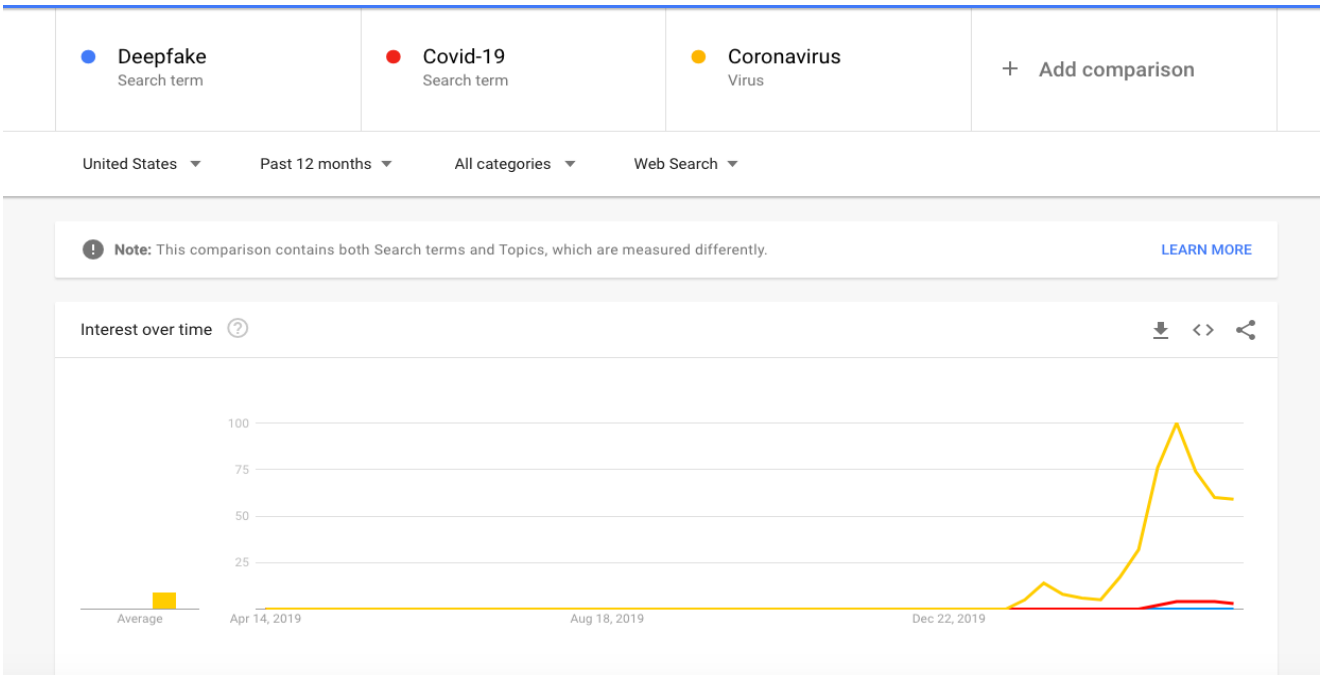


Figure 4. Deepfake vs Coronavirus vs Covid-19 search Trend (Google Trends, 2020).

Accounting for the fact that users can also search for Coronavirus we see that all of the previous hope that was provided was indeed false. However, this comparison still lacks quantative information to do any analysis on. To fix this we will look at how popular deepfakes were in comparison to Facebook as Facebook's data on user traffic can be leveraged to come up with some rough numbers for the number of users learning about deepfakes. The results are below:


Figure 5. Deepfake vs Facebook search Trend (Google Trends, 2020).

To give the biggest possible estimated value for deepfakes searches and therefore users learning about deepfakes we can do the following mathematics. Facebook had on average about 2.196 billion monthly active users in Q1 2018 (Clement, 2020). Let us assume that every single one of those users logged into Facebook by using a Google search during February 2018, a month where Facebook's is index around 53. Due to the fact that these indexes are proportional to each other we can set index value of 53 to be about 2.196 billion searches. By generously rounding up the index value of deepfakes to one and by operating under these very faulty assumptions, we get a value of about 41 million searches for deepfakes for the month of February. Let us assume that every single one of these searches is a brand-new user and that this trend continues for two years (Which it does not as seen from figure 1). With that we arrive at around 980 million users knowing about deepfakes. However, this value, although very large, does not cover the estimated 3.8 billion social media users in 2020(Carter, 2020). As a result, this lack of awareness about deepfakes makes many social media users susceptible to be fooled by

the technology and this only gives more reason for why social media platforms need to change their current content management methodologies in response to deepfakes.

*Policy Document Analysis*

Now that the user's position relative to the problem of deepfakes has been more fleshed out, I wanted to see the three platforms positions on the issue. Below is a table of all three platforms formal stances on deepfakes:

| **Social Media Platform** | **Policy pertaining to deepfakes** |
|---|---|
| Reddit | "Content is prohibited if it Impersonates someone in a misleading or deceptive manner". ("Reddit Content Policy", n.d). |
| Twitter | "Synthetic and manipulated media: You may not deceptively share synthetic or manipulated media that are likely to cause harm. In addition, we may label Tweets containing synthetic and manipulated media to help people understand their authenticity and to provide additional context." ("The Twitter Rules", n.d). |
| Facebook | "Media, including image, audio, or video, can be edited in a variety of ways. In many cases, these changes are benign, like a filter effect on a photo. In other cases, the manipulation isn't apparent and could mislead, particularly in the case of video content. We aim to remove this category of manipulated media when the criteria laid out below have been met.<br>In addition, we will continue to invest in partnerships (including with journalists, academics and independent fact-checkers) to help us reduce the distribution of false news and misinformation, as well as to better inform people about the content they encounter online.<br>Do not post: Video that has been edited or synthesized, beyond adjustments for clarity or quality, in ways that are not apparent to an average person, and would likely mislead an average person to believe that a subject of the video said words that they did not say AND is the product of artificial intelligence or machine learning, including deep learning techniques (e.g., a technical deepfake), that merges, combines, replaces, and/or superimposes content onto a video, creating a video that appears authentic. This policy does not extend to content that is parody or satire or is edited to omit words that were said or change the order of words that were said." ("Community Standards",n.d). |

Table 1. Social Media Content Policies (Ghatti, 2020)**.**

It can be seen that under the current policy of all three social media platforms, Reddit, Twitter, and Facebook, there is no allowance for deepfakes. In fact, Facebook's content policy explicitly bans deepfakes on their site that are not parody or satire, a move that the company made in January of this year (Bickert, 2020). This document analysis serves to further answer the why part of the research question. The reason why social media platforms must respond to

deepfake technologies by changing their current content management methodologies is due to the fact that their content policies do not, explicitly or implicitly, allow for it to be on their platform.

*Current Automated Solutions*

To uphold these policies social media platforms can look to automated solutions. Due to the fact that deepfake technology is relatively new, there are not many automated solutions that exists to detect deepfake videos. Many of the current automated solutions that do exists look at the warping effects that are left in the deepfakes during production. One example of such a solution is from Guerra and Delp (2018) who designed a deepfake detection system using a Convolutional LTSM. A Convolutional LTSM is a neural network composed of a Convolution Neural Network (CNN) and Long Short-Term Memory Network (LTSM). A CNN is an algorithm that operates much like the visual cortex of the human brain. It takes an image and assigns different levels of importance to different features of the image. Using these varying levels of importance, the CNN is able to differentiate and classify images (Saha,2018). A LTSM, on the other hand, is a network that allows for the learning of long-term patterns and dependencies between data (Bronwlee, 2017). Using these two technologies, the system designed by Guerra and Delp (2018) exploits some of the flaws that are apparent with deep fakes. One such flaw is the inconsistencies that exist between the swapped faces and the rest of the scene in the background. Guerra and Delp (2018) found that their system could predict whether a video was a deepfake or an original video with a 96 percent accuracy rate.

**Discussion**

Although a 96 percent accuracy rate is excellent, all of the techniques that are used in Guerra and Delp's identification systems will soon be obsolete due to the fact that deepfakes

constantly get better by correcting any flaws that are found. This aspect of deepfake technology will also play an impact on future research that is done on detection tools as the results of this research will have to be kept secret in order to stay effective.

The constant iterations of improvement made by deepfake technology towards realism is one half of the equation as to why deepfakes pose such a large problem to social media platforms. The other half has to do with the perception of audio and video being "incorruptible" forms of evidence. Going back to Actor Network Theory (ANT), the historical difficulty of proctoring realistic video and audio artifacts has led to these technologies prescribing a behavior of faith and trust from the viewer (Latour, 1992). Video and audio that looks "real" is never questioned and is taken at face value. Deepfakes can be used to exploit this behavior, shifting power away from social media platforms and towards nefarious agents who can exploit deepfakes to essentially manufacture information at will on these platforms.

In order to avoid such a grim future, I propose automated detection systems be utilized by social media platforms on top of their current moderation systems to combat against deepfakes. Such a configuration could result in a system in which the deepfake detectors flag content before the content is sent to the human moderators. Artificial intelligence systems may be good at detecting what content is a deepfake, but these systems would struggle when trying to make the ambiguous decision of what is appropriate for the platform. Going back to Latour's Actor Network Theory (ANT), specifically the delegation of labor, artificial intelligence systems would not be reliable enough to justify completely delegating duties from human moderators to these automated systems. However, by having a moderation infrastructure that uses both human moderators and automated detection systems, a combined program of action is created that can effectively combat deepfakes. Specifically, the combined program of action of "detect and mark

deepfakes" of the automated system and "moderate content (including deepfakes)" of the content moderators would result in a program of action that moderates content while also effectively battling nefarious deepfakes (Latour, 1992).

**Limitations**

Although my personal program of action was to find values for how many internet users know about deepfakes, the best I could do was find a rough estimate. As a result, the biggest limitation in my research is the lack of actual values for the section estimating how many internet users are aware of deepfakes. Due to the lack of actual values, the current research in this area that I have presented operates under extremely unrealistic assumptions in order to engineer a value.

**Lessons Learned**

Fixing the above limitation would be the first action that I took if I did the research again. I would have done this by supplementing the section about deepfake awareness with survey results from my local community about their knowledge on deepfakes. The survey could also consider demographics such as education level and occupation. There is a lot that I could have learned by conducting that survey. Almost as much as I have learned from this research. Perhaps the most prominent lesson that I have learned from this research that I will continue to implement in my future in engineering is how interlaced technology and society are. From a technical perspective, deepfakes are nothing but an application of machine learning techniques. However, this application has a wide variety of social implications encompassing issues such as misinformation and censorship. I have learned that I must be aware of the implications of my work. I have also learned to be careful when levering the natural tendencies of users in technology. This is a large focus in computer science, especially in human-computer interaction,

where the goal is to make it seem like the technology isn't even there. However, after seeing how deepfakes leverage viewer's susceptibility to believe video and audio in order to spread misinformation, I now understand that sometimes it is important to make the user aware of the presence of technology.

**Conclusion**

Deepfake technology is an approaching threat and action needs to be taken now in order to ensure that nefarious agents cannot literally manufacture information for consumption. Social media platforms serve to be the primary tool that can be utilized to spread misinformation with deepfakes and as a result these platforms must take steps in order to ensure that their platform is not hijacked. When deepfakes inevitably do arrive, social media platforms will have to deal with issues such as censorship, what is allowed on the platform, and freedom of speech rights. These issues will undoubtedly be complex and finding a satisfactory answer will be difficult. However, for now, being able to tell what's real and what's not is a great start.

**References**

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives, 31*(2), 211–236. doi: 10.3386/w23089

Bickert, M. (2020, January 14). Enforcing Against Manipulated Media. Retrieved from https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/

Brownlee, J. (2019, August 14). A Gentle Introduction to Long Short-Term Memory Networks

by the Experts. Retrieved from https://machinelearningmastery.com/gentle-introduction-long-short-term-memory-networks-experts/

Carter, J. (2020, February 28). Search engine marketing statistics 2020. Retrieved from https://www.smartinsights.com/search-engine-marketing/search-engine-statistics/

Clement, J. (2020, January 30). Facebook users worldwide 2019. Retrieved from https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Dack, S. (2019, July 11). Deep Fakes, Fake News, and What Comes Next. Retrieved from https://jsis.washington.edu/news/deep-fakes-fake-news-and-what-comes-next/.

Desai, U. (2018, April 29). Keep Calm and train a GAN. Pitfalls and Tips on training Generative Adversarial Networks. Retrieved from https://medium.com/@utk.is.here/keep-calm-and-train-a-gan-pitfalls-and-tips-on-training-generative-adversarial-networks-edd529764aa9

Dwoskin, Elizabeth J. W. (2019, July 25). Content moderators at YouTube, Facebook and Twitter see the worst of the web - and suffer silently. Retrieved from https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price/.

Facebook (2020) Community Standards. (2020) Retrieved from

https://www.facebook.com/communitystandards/manipulated_media.

Gillespie, T. (2018). Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press, New Haven CT.

Google Trends. (2020) Retrieved from https://trends.google.com/trends/

Guera, D. J., & Delp, E. J. (2018). Deepfake Video Detection Using Recurrent Neural Networks. 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). doi: 10.1109/avss.2018.8639163

Latour, B. (1992). Where are the missing masses? Sociology of a few mundane artefacts. In Bijker, W. E., and Law, J. (eds.), *Shaping Technology-Building Society: Studies in Sociotechnical Change*, MIT Press, Cambridge, Mass.

Reddit (2020) Reddit Content Policy Retrieved from https://www.redditinc.com/policies/content-policy.

Robertson, A. (2018, February 7). Reddit bans 'deepfakes' AI porn communities. Retrieved from https://www.theverge.com/2018/2/7/16982046/reddit-deepfakes-ai-celebrity-face-swap-porn-community-ban

Rogers, S. (2016, July 1). What is Google Trends data - and what does it mean? Retrieved from https://medium.com/google-news-lab/what-is-google-trends-data-and-what-does-it-mean-b48f07342ee8

Saha, S. (2018, December 17). A Comprehensive Guide to Convolutional Neural Networks - the ELI5 way. Retrieved from https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53.

Search Engine Market Share in 2019. (2020). Retrieved from https://www.oberlo.com/statistics/search-engine-market-share

Scola, N. (2019, May 24). Facebook on fake Pelosi video: Being 'false' isn't enough for removal. Retrieved from https://www.politico.com/story/2019/05/24/facebook-fake-pelosi-video-1472413

Shen, T., Liu, R., Bai, J., & Li, Z. (2018). “Deep Fakes” using Generative Adversarial Networks (GAN). Retrieved from http://noiselab.ucsd.edu/ECE228_2018/Reports/Report16.pdf.

Twitter (2020) The Twitter Rules. Retrieved from https://help.twitter.com/en/rules-and-policies/twitter-rules.

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151. doi: 10.1126/science.aap9559

Zhao, Z., Zhao, J., Sano, Y. et al. Fake news propagates differently from real news even at early stages of spreading. EPJ Data Sci. 9, 7 (2020). https://doi.org/10.1140/epjds/s13688-020-00224-z