DEEPFAKES' DEEP SCARS ON DEMOCRACY

A Research Paper submitted to the Department of Engineering and Society
Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

By

Nathan Williams

March 30, 2023

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.

ADVISOR
Catherine D. Baritaud, Department of Engineering and Society

TANGIBLE AND INTANGIBLE THREATS OF DEEPFAKES

Perhaps one of humanity's crowning achievements is the development of technology for recording and exchanging information. From cave paintings to encyclopedias, the written word has served as the primary method of recording history. However, written text is only as trustworthy as its author and with the invention of cameras and videos came a new and more trustworthy standard of media, one that was much harder to tamper with or fabricate. Now, in the early 21st century, incredible progress in computer science and machine learning threatens the credibility of all forms of media.

DEFINING A DEEPFAKE

Brooks et al. (n.d.) define deepfakes as a subset of synthetic media that uses machine learning to create "believable, realistic videos, pictures, audio, and texts of events which never happened" (p.3). A person looking to create a deepfake needs to have a few key ingredients. First a target is needed to be the subject of the deepfake, for videos, this would be the person whose face will be "pasted" onto a new body. More video data of the target's face and videos at different angles help the deepfake model learn the target's face more accurately (Falis, 2020, p. 623). This is why the majority of deepfakes target celebrities or politicians who would have plenty of videos available publicly (Mustak et al., 2023). The host in a deepfake refers to the video which the target's face is inserted into. While the host can be any person, making a convincing deepfake requires a host of similar body type as well as matching the lighting in the target video (Brooks et al., n.d.). Finally, all good deepfakes require either significant computing power and/or time. The machine learning models for deepfakes are intensive, and benefit greatly from strong graphic processing units. In lieu of a strong GPU, a model can be trained with lower computing power at the cost of significantly increased training time (Mustak et al. 2023).

THE SOCIETAL DANGERS OF DEEPFAKES

Spreading misinformation

While many political deepfakes are obvious jokes and not particularly well made, they can still influence society's perception of political candidates. One notable instance is when Nancy Pelosi, speaker of the United States House of Representatives, is portrayed giving an interview while slurring her words and appearing intoxicated. This deepfake was then reposted by President Donald Trump with the caption: "PELOSI STAMMERS THORUGH NEWS CONFERENCE" (Mervosh, 2019, p. 1). Even though the video was eventually debunked, Trump never deleted the tweet, and it was likely seen by a large amount of his followers. Another way deepfaked misinformation could be used to threaten democracy is generating blackmail. By fabricating video or audio where a politician is seen in a compromising scenario, such as a pornographic video or committing a hate crime, a malicious actor could then threaten the politician with releasing the video to their friends, family, or the public, ruining their career in the process.

Regina Rini (2022), a professor and research chair of Philosophy of Moral and Social Cognition at the University of Wisconsin highlights another potential danger of deepfakes in the form of panoptic gaslighting where "a person's memory and identity are undermined by a myriad of systemically targeted fabrications" (143). While definitely the least likely of threats, panoptic gaslighting has the potential for the most severe impact. Being able to convince a politician that they took a stance or spoke on an issue they never had before would give a malicious actor a devastating level of control over a politician.

Liar's dividend: Destroying a shared reality

Another complicating factor to the influence of deepfakes on politics is the concept of liar's dividend. Maria Pawlec (2022), a professor and member of the International Center for Ethics in the Sciences and Humanities at the University of Tübingen defines liar's dividend in the context of deepfakes as "the opportunity for individuals criticized for certain statements or actions to simply deny the truthfulness of incriminating evidence by referencing the existence of deepfakes" (p. 12). An example of this in the United States is when President Joe Biden released a video updating the American people on the January 6th Capitol riots. In response to this video his political opponents simply dismissed the announcement as a deepfake. It was even publicly declared fake by U.S. news outlets Newsmax and One America News (Horton & Sardarizadeh, 2022). A more extreme example of deepfakes stoking liar's dividend is when the health of Gabonese president Ali Bongo was called into question by political opponents in 2019, some even claiming he was dead. President Bongo released a video to prove he was healthy and capable of leading, but it was labeled a deepfake by his opposition and soon after they attempted and failed to throw a military coup (Smith & Mansted, 2020, p.13).

Similar to liar's dividend but at a more generalized level, is the erosion of a shared reality. Greg Zachary (2020), a professor and member of the International Center for Ethics in the Sciences and Humanities at the University of Tübingen, explains that when media, especially news and journalism, loses all credibility in the public eye a society's sense of shared reality deteriorates (p. 110). When every member of society has their own reality and facts, there is no basis for any sort of productive debate, therefore increasing the spread and belief of misinformation.

RESPONSES TO DEEPFAKES AND MISINFORMATION

Since deepfakes already have the potential to cause significant harm to societies, it is important that societies develop an effective response to address and mitigate the dangers of deepfakes, especially within social media. This creates a complex and interconnected web involving national governments, international social media platforms, and general populations all with unique interests and limitations. Actor Network Theory (ANT) is especially useful for analyzing these relationships between human, governmental, and organizational actors since ANT is "able to focus on the numerous moments of translation as they are enacted in the process of building the sociotechnical" (Cressman, 2009, p. 9). The policies and legislation of deepfake regulations are all moments of translation that ANT excels in unravelling and organizing. The remainder of this thesis will use Actor Network Theory in a case study comparison of different approaches to prevent or mitigate the harms of deepfakes. The comparison of these actor networks will provide insight for regulating deepfake usage effectively by examining the benefits and drawbacks of these unique networks. Coupled with the state-of-the-art technical report portion, this thesis aims to provide insight into combatting synthetic media misinformation through technological and social avenues.

TRACING THE JOURNEY OF A DEEPFAKE

To first understand the actor networks that surround stopping harmful deepfakes, it is necessary to understand how those deepfakes wreak havoc in the first place. The handoff model shown in Figure 2 can be used to visualize the journey of a political deepfake, providing a more linear framework for understanding when and where actors are able to stop the consequences of a deepfake. This model shows how malicious actors can use social media to either blackmail

politicians directly or to unknowingly recruit media consumers to spread their misinformation out of ignorance.

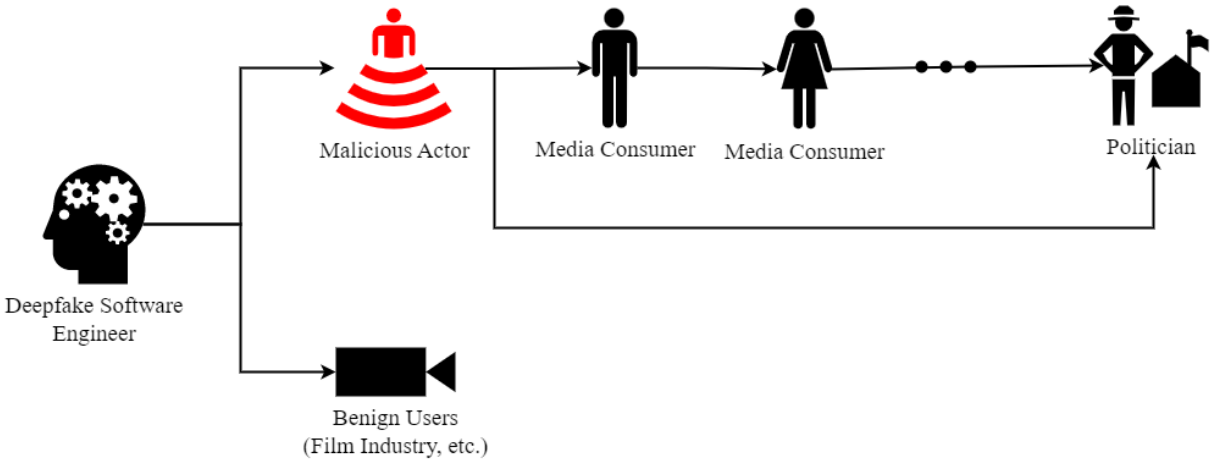

Figure 1: Handoff Model for Deepfake Technology. This figure demonstrates the journey of deepfake technology as it spreads throughout society. (Adapted by Williams (2022) from W. Carlson & C. Baritaud, class handout, 2009).

To stop or mitigate the handoff of deepfaked misinformation, the main tools at a society's disposal are counter-technology, platform-based solutions, and governmental legislation.

FIGHTING FIRE WITH FIRE: COUNTER-TECHNOLOGY

As the quality of deepfakes becomes increasingly realistic with technological advancements, humans are less likely to identify deepfakes with their judgement alone. As such, the ideal way to combat deepfaked misinformation is to utilize a technology that would be able to accurately identify whether certain media is deepfaked or not. When a deepfake is rendered, there are no flags in its file data that would indicate it has been tampered with, it simply exists as a collection of frames played quickly enough to give the notion of movement (Groh et al., 2021, p. 3). Therefore, a simple file scanner would not be enough to detect tampering, instead it is necessary to use the very same technology that created the deepfake to detect it: machine learning models trained to classify media as pristine or deepfaked.

With this technology, social media corporations such as Facebook, Twitter, etc. would be able to automatically classify content as it is uploaded and either remove it or label it accordingly. This would essentially counter any attempts at spreading misinformation through deepfakes. However, while models exist for detecting deepfakes, they are not nearly accurate enough to be acceptable for commercial use. In 2019 tech giants such as Meta and Microsoft, as well as multiple prestigious universities like Cornell and M.I.T. came together to host the Deepfake Detection Challenge (DFDC). Competing for a prize total of $10 million, both amateurs and experts were invited to create a model for detecting deepfaked videos. The winning model scored an accuracy of 65.18% against videos not used for model training, which while impressive for the challenge is not hopeful for widespread use ("Deepfake Detection Challenge Results: An open initiative to advance AI", 2020).

Since the DFDC in 2019, deepfake detection models have been constantly improving with some models like Shohel Rana and Andrew Sung's (2020) DeepfakeStack, reporting accuracy rates in the 90^{th} percentile; however, it should be noted that this model used the same videos for training and testing data and has no reported accuracy on videos it has not been trained on (p. 74). Without reliable deepfake detection software, deepfake based misinformation must be solved through social methods.

PLATFORM POLICY BASED SOLUTIONS

Without the technology to completely shut down deepfaked misinformation, another path to mitigating the damage of deepfakes is having social media platforms implement policies that target misinformation. Social media giant Meta states in their policy on synthetic media that content will be removed if it "has been edited or synthesized – beyond adjustments for clarity or quality – in ways that aren't apparent to an average person" (Bickert, 2020). Similarly, Twitter's

policy is that customers "may not share synthetic, manipulated, or out-of-context media that may deceive or confuse people and lead to harm" ("Synthetic and manipulated media policy", n.d.). Figure 3 shows twitters guidelines for assessing and responding to synthetic media it finds.

Is the media significantly and deceptively altered or fabricated?	Is the media shared in a deceptive manner?	Is the content likely to impact public safety or cause serious harm?	
✓	✗	✗	Content **may** be labeled
✓	✗	✓	Content is **likely** to be labeled, or **may** be removed.
✓	✓	✗	Content is **likely** to be labeled.
✓	✓	✓	Content is **very likely** to be removed.

Figure 2: Table of Twitter's Synthesized Content Removal Policy. This figure shows the considerations in labeling and removing deepfaked content (Roth, Y. & Achuthan, A. 2020).

It is worth mentioning that both platforms have stipulations in their policies that excuse synthetic media created for the purpose of satire which leaves a significant grey area as to what content qualifies as satire. Furthermore, Meta reports that the vast majority of content it removes is chosen for removal by their artificial intelligence ("How technology detects violations", 2022). As deepfakes become more prevalent on social media, and with deepfake detection classifiers barely breaking 60% accuracy, it is unreasonable to assume that an A.I. would be able to accurately detect deepfaked media in the near future, much less understand the nuances of satire.

While the majority of removed media on social media platforms is detected automatically, most of the big-name social media platforms employ 3rd party fact checkers to varying degrees. Instagram states that they usually reserve the use of fact checkers such as the Associated Press or Rueters for trends and other viral-esqe phenomenon that are more likely to have widespread impact ("How Meta prioritizes content for review", 2022). Furthermore,

compared to using company owned automated software for policing misinformation, 3rd party fact checkers are extremely expensive for platforms to hire. Meta reports having spent $100 million on fact-checking efforts since 2016, a costly endeavor that any investor would likely want to avoid if possible ("Meta's Investments in Fact-Checking", 2022). With social media platforms missing adequate detection technology, and leaving their flagging/removal policies vague, other options for combatting misinformation must be explored.

GOVERNMENTAL LEGISLATION

Many countries have laws specific to how social media platforms can operate within their borders, this extends to policies related to policing misinformation and synthetic media such as deepfakes. By using Actor Network Theory to analyze the current systems in the United States, China, and Australia a policy maker could discern the strengths and weaknesses of these countries' systems and gain insight into building a better system.

United States of America

In the United States, there are some states that have laws targeting synthetic media specifically, such as in California where it is illegal to create a deepfake "with the intent to injure a candidate's reputation or deceive a voter into voting for or against the candidate" (O'Donnell, 2021, p. 715). However, at the federal level there are no laws specific to synthetic media. Instead, The Communications Decency Act of 1996 (CDA) details responsibility for misinformation on social media platforms, and in essence "protects online service providers from legal liability stemming from content created by the users of their services" (Zachary, 2020, p. 109). In this case, responsibility for damages caused by deepfakes and other misinformation falls upon the original poster. The problem with this approach is that due to I.P. address maskers and

virtual private networks anyone intent on spreading misinformation could easily cover their tracks and become impossible to find (O'Donnell, 2021, p. 718).

In response to the rise in misinformation, Congress recently introduced the Educating Against Misinformation and Disinformation Act (EAMDA) which proposes mass citizen education on recognizing misinformation including deepfakes (Educating Against Misinformation and Disinformation Act, 2022). None of the proposed education plans in the EAMDA were mandatory or included in public school curriculum, as such, it is unclear how far reaching the bill will be if passed. Regardless, the primary issue remains; U.S. legislation places the onus of fighting misinformation on the victims of deepfaked media. These laws take the responsibility of handling misinformation and give it to social media platforms without enforcing consequences for irresponsible management. Furthermore, these platforms as privatized corporations may not always have the best interests of the public at heart, especially when 3rd party fact checkers are so expensive. For example, in the aforementioned instance of a deepfake showing an intoxicated Nancy Pelosi, Facebook simply marked the video as "partly false" and Twitter did not address the video at all (Denham, 2020).

If policy makers considered ANT when discussing legislation surrounding synthetic media, they would see that the current system is ineffective in combatting misinformation. Figure 3 visualizes the American actor network for addressing synthetic media misinformation and shows the breakdown in moments of translation.

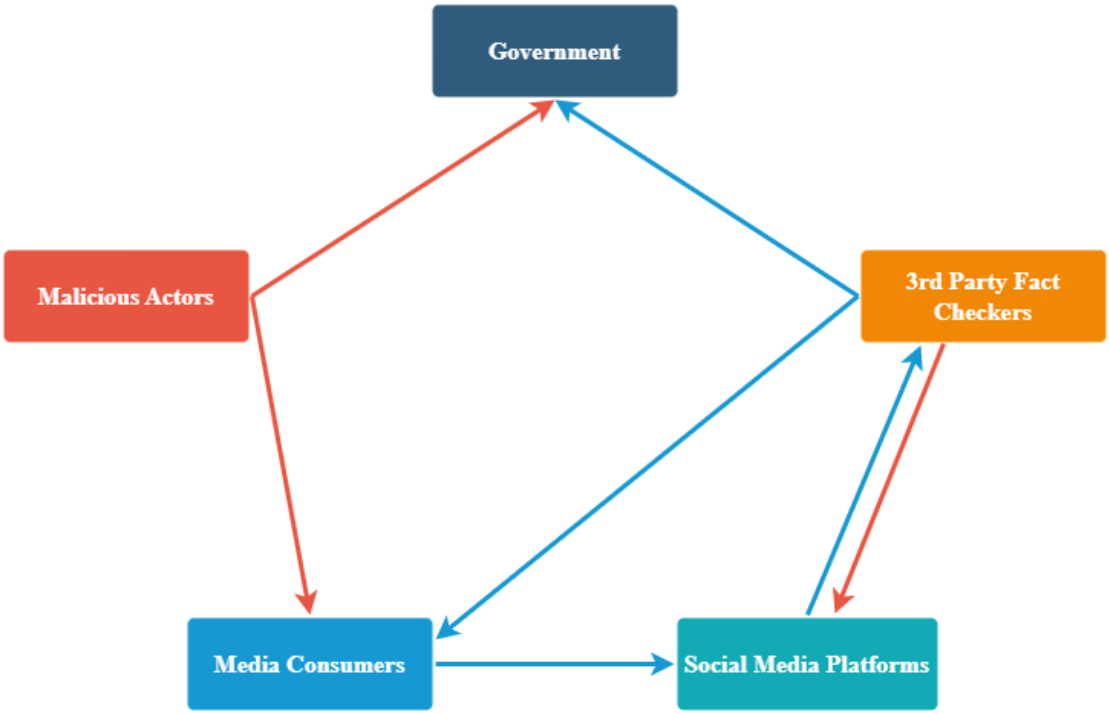

Figure 3: Actor-Network Theory analysis of the United States approach to regulating misinformation. The relationships between legislative, provider, and consumer actors determine the quality of deepfake regulation. (Adapted by Williams (2023) from Callon (1984, 1987) & Latour (1987)).

While media consumers continue to support social media platforms, the only actors that offer protection from misinformation to consumers and governmental officials are 3rd party fact checkers. This protection is not mandatory, causing many social media platforms to hesitate in hiring fact checkers when considering the monetary costs. Actor Network Theory shows that without the technology to enforce preventative screening for deepfakes, and with legislation that holds no one responsible for the aftermath, the United States government is failing to address the issue of synthetic media and misinformation.

China

In China the social media platforms that westerners are used to such as Facebook and Twitter are banned, in their place the app Weibo which has 8 times more users than Twitter dominates the social media landscape (Hine & Floridi, 2022, p. 608). In January of 2022 the

Chinese government passed a set of laws that specifically target deepfakes, these laws require that all deepfaked material be watermarked and incurs steep punishments, including jail time, for spreading "rumors" with synthetic media (Hine & Floridi, 2022, p. 609). Yang et al. (2012) from the Shandong Provincial Key Laboratory of Software Engineering explain that enforcing regulations such as these is possible because the government has mandated that Weibo use automatic "rumor-busting" software, where the government decides what qualifies as rumor or not (p. 2). Furthermore, Weibo requires users to sign up with their real names verified by the government and does not allow for anonymous posting which ensures that those who break regulations can be held accountable (Lee & Liu, 2016).

In essence, the Chinese government's hardline approach to policing deepfaked media and misinformation is the opposite of the United States' laissez faire strategy. Unfortunately, it is difficult to discern the effectiveness of China's new deepfake legislation since Weibo does not release data on how many posts it removes under the pretense of misinformation. Nonetheless, it is useful to look at the Chinese legislation through ANT to discuss the ethical consequences of deepfake and misinformation legislation. The main difference between other deepfake regulation actor networks and China's is that China has removed 3rd party fact checkers as an actor as shown in Figure 4.

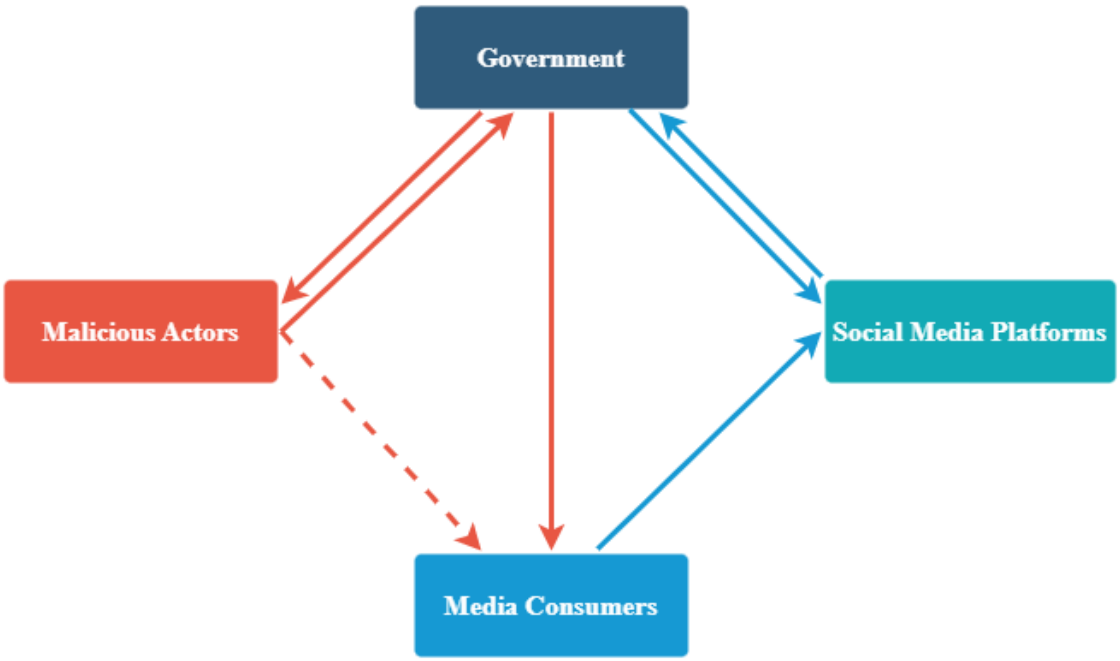

Figure 4: Actor-Network Theory analysis of China's approach to regulating misinformation. A network that substitutes government oversight for 3rd party fact checkers. (Adapted by Williams (2023) from Callon (1984, 1987) & Latour (1987)).

In the place of fact checkers, the government decides what content is removed from platforms. So long as social media platforms comply with providing user data to the government, they are allowed to operate in Chinese webspace (Bamman et al., 2012). However, while malicious actors are still able to post deepfaked misinformation, the automatic rumor busting algorithms employed by Weibo would likely remove the post before it could gain significant traction among media consumers offering them and the government protections from misinformation (Yang et al., 2012). Furthermore, the government's ability to identify malicious actors along with the severe penalties for spreading misinformation would likely deter many malicious actors from posting in the first place.

Despite the strengths of accountability and rapid enforcement in the Chinese actor network there is a significant consequence of removing 3rd party fact checkers as an actor. The consequence of placing the responsibility of media sanitization on the government paves the way

for widespread censorship. Emmie Hine (2022) from the Digital Governance Group in Oxford, and Luciano Floridi (2022) from the Oxford Internet Institute explain how the new provisions show "a prescient understanding of how new technologies could threaten social stability and thus the regime's power" in China (p. 610). Furthermore, based off similar past internet regulations it seems that the focus of the legislation is to limit outspoken opposition more than prevent misinformation from spreading (Hine & Floridi, 2022). As such, the Chinese response fails to properly the address the deepfaked misinformation crisis by sacrificing privacy and free speech for the sake of efficient enforcement.

Australia

Australia has taken a more proactive approach than the United States but is not nearly as overbearing as Chinese legislation. As of now, Australia has similar laws to the United States that absolve social media companies of responsibility for misinformation content on their platforms (Hurcombe & Meese, 2022). However, Australia has taken further steps by forming the Digital Industry Group Inc. (DIGI), which is a not-for-profit industry association. DIGI began when the Australian Communications and Media Authority (ACMA) invited tech and media giants like Google, Meta, TikTok, etc. to come together and create the Code of Practice on Mis- and Disinformation, which outlined industry standards for dealing with misinformation including synthetic media (Hurcombe, E., Meese, J. 2022). These standards include a commitment to provide periodic transparency reports that will show users how and why content is removed. Furthermore, the code suggests that platforms label/remove known false content or if content cannot be verified by a 3rd party, the platform must provide a reasonable means for the user to check themselves (Digital Industry Group Incorporated [DIGI], 2021). These codes have already proven to be an improvement to the previous system as Facebook (2021) reports

removing an extra 110,000 instances of misinformation from Australia and that 6.2 million Australians visited their transparency report. Considering these promising results, a new actor network can be constructed in Figure 5 to represent the Australian approach to combatting deepfaked misinformation.

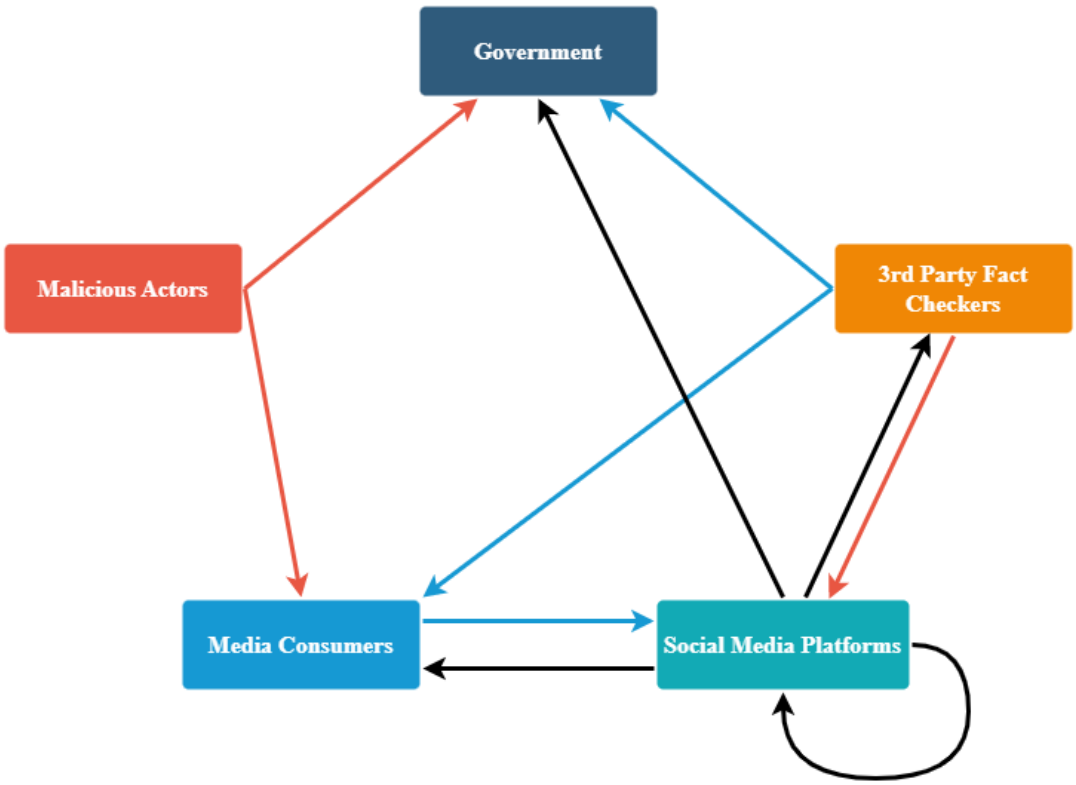

Figure 5: Actor-Network Theory analysis of the Australian approach to regulating misinformation. A complex network that includes optional self-regulation for social media platforms. (Adapted by Williams (2023) from Callon (1984, 1987) & Latour (1987)).

The implementation of DIGI expands the actor network so that social media platforms become responsible for holding each other accountable. Furthermore, the code of practice has been shown to remove more misinformation and educate media consumers through transparency reports, which helps to protect both media consumers and government officials from misinformation.

The main issue with this actor network is that currently DIGI is an opt-in organization and social media platforms are not required to join it, and so its benefits are not guaranteed.

However, in January, 2023 the Minister for Communications Hon Michelle Rowland (2023) announced that the ACMA will be granted new powers to enable them to "register an enforceable industry code" and "be responsible for the content they host and promote", essentially making the DIGI Code of Practice mandatory for all media platforms that wish to operate in Australia and creating punitive measures for platforms that break the code ("New ACMA powers to combat harmful online misinformation and disinformation", 2023). In response to this announcement the DIGI Managing Director Sunita Bose "broadly welcomes the Government's announcement … and look forward to engaging with the details during public consultation" ("DIGI Welcomes The Government Providing ACMA With Oversight Powers Over Misinformation", 2023). ACMA has taken proactive measures in ensuring that the responsibility of combating synthetic media and other misinformation does not fall on the everyday citizen, while also ensuring that social media platforms are not given complete free reign in policing misinformation. The promising increase in misinformation removal and transparency viewership indicates that Australia is on the right track in addressing deepfakes and misinformation with government regulation.

IMPROVING ON EXISTING ACTOR NETWORKS

After analyzing the actor networks of the United States, China, and Australia it is necessary to propose a new model that highlights the strengths of existing networks while attempting to allay their weaknesses. At the heart of this proposed model is a code of practice, similar to the one in Australia, that would require social media platforms to remove, or flag known misinformation, as well as provide transparency reports for consumers. DIGI's success in Australia shows that platforms taking these steps can be effective in mitigating the spread of misinformation.

Where the proposed network differs from Australia's model is in how the code is enforced. Learning from America's inaction and drawing from China's use of legislation, the proposed model's code of practice would be codified into law. Thus, making compliance mandatory for any social media platforms that wish to operate in the country and guarantee that the benefits of a code would be realized. Punitive measures, such as fines, for failing to remove known misinformation, or refusing to produce transparency reports would provide incentive for social media platforms to adhere to the agreed upon code of practice. Figure 6 provides a visualization of how the proposed actor network would operate.

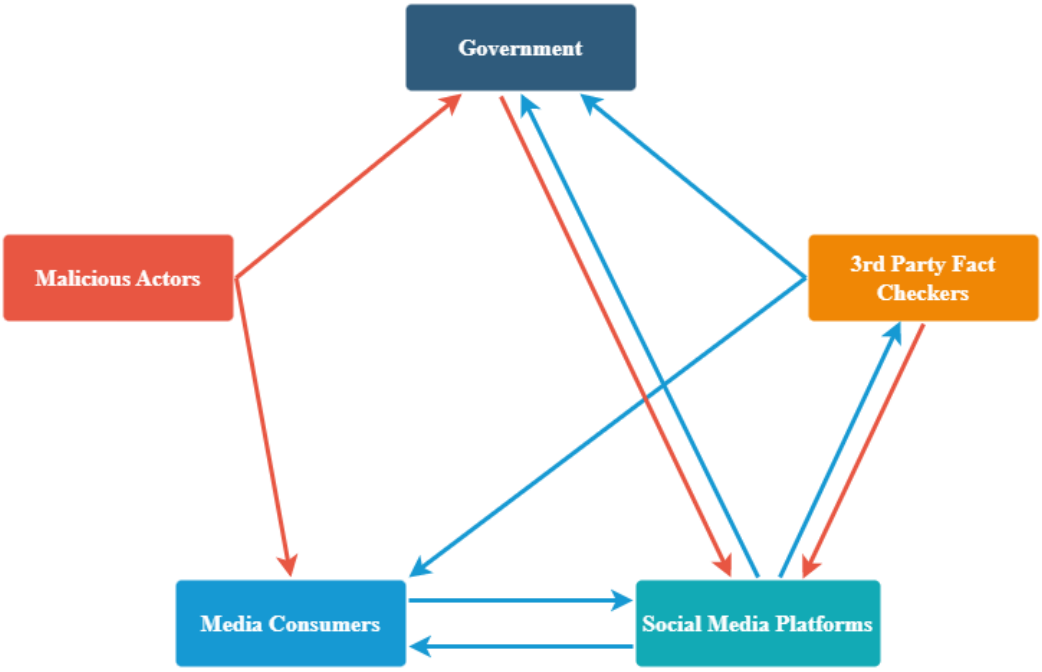

Figure 6: Actor-Network Theory analysis of the proposed approach to regulating misinformation. A network showcasing a combination of a code of practice and punitive legislation. (Adapted by Williams (2023) from Callon (1984, 1987) & Latour (1987)).

Another point of note with the proposed model is that by having social media platforms responsible for identifying misinformation with the help of 3rd party fact checkers the government would not be able to directly censor content on the platforms. Furthermore, public

transparency reports would make it clear how and why content is flagged/removed so any unjust removals could be publicly scrutinized.

To understand how the proposed actor network would affect the lifecycle of deepfaked misinformation the updated handoff model in Figure 7 outlines how social media platforms would engage with the distribution of synthetic media. While there is currently no surefire way for platforms to identify and stop the spread of deepfakes before they reach media consumers or politicians, it is certainly possible for platforms to invest resources into identifying and flagging or removing deepfaked content as soon as possible, thus mitigating the spread of malicious deepfakes.

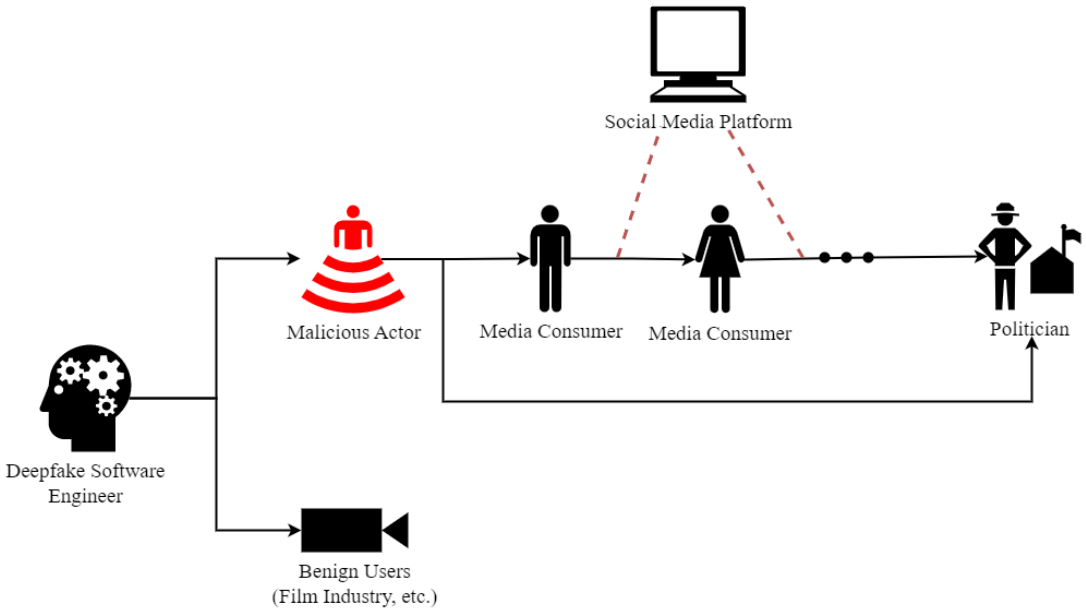

Figure 7: Proposed Handoff Model for Deepfake Technology. This figure how the spread of deepfaked misinformation would be addressed in the proposed actor network. (Adapted by Williams (2022) from W. Carlson & C. Baritaud, class handout, 2009).

THE FUTURE IMPLICATIONS OF SYNTHETIC MEDIA MISINFORMATION REGULATION

In the future, it is entirely possible that advancements in deepfake detection A.I. models could allow social media platforms to automatically and instantly flag or remove deepfaked

content accurately. If and when this technology is available it is likely that many social media platforms will implement it of their own accord. By severing the distribution chain between a malicious actor and media consumer, this would all but completely negate the societal danger deepfakes pose. Until such technology exists, the next best option is to mitigate the harm of deepfakes as they spread and have regulations in place that will help to identify and flag deepfaked misinformation as soon as possible.

In the case of the United States, the future of legislation surrounding deepfake laws is unclear. The Educating Against Misinformation and Disinformation Act has yet to be brought up for discussion in Congress and is likely months if not years away from potentially being enacted (Educating Against Misinformation and Disinformation Act, 2022). Until then, the social media landscape will continue to be run unchecked by Meta, Twitter, etc. and any sort of action taken against malicious deepfakes will be up to their discretion and interests.

On the other hand, the new provisions on synthetic media in China is another addition in a string of laws designed to allow the government significant oversight and influence in the country's cyber space such as the real-name verification and rumor-busting laws of the early 2010's. Therefore, it is likely that in the future, the legislation on deepfakes will remain as is or become even stricter, either through regulating the production of deepfake technology or imposing harsher measures on those who post deepfakes.

Australia has already seen positive results in preventing the spread of misinformation through its implementation of DIGI and a code of practice. Moving forward, if the ACMA decides to make DIGI membership mandatory for all social media platforms it is not unreasonable to assume that the benefits in content integrity seen in current DIGI members will spread to other platforms. Furthermore, Edward Hurcombe (2022) from the Queensland

University of Technology Digital Media Research Centre suggests that making platforms responsible for the content they host would encourage commercial research into deepfake detection technologies which would provide a much more effective and reliable long-term solution (p. 301).

For all countries looking to combat synthetic media spreading misinformation it is helpful to contextualize the issue with Actor Network Theory to help identify the responsibilities and relations of national governments, international social media platforms, and general populations. Through this case study comparison, the most effective actor networks for regulating misinformation and by extension, deepfakes, was found to involve governments working closely with the providers of technology to draft legislation that protects corporate and user interests but stills holds platforms accountable for the content they host. This approach to regulating deepfakes should be considered in the future by organizations and countries looking to mitigate the harm of synthetic media.

REFERENCES

Australian code of practice on disinformation and misinformation - digi. DIGI. (2021, February 22). Retrieved March 23, 2023, from https://digi.org.au/wp-content/uploads/2021/02/Australian-Code-of-Practice-on-Disinformation-and-Misinformation-FINAL-PDF-Feb-22-2021.pdf

Bamman, D., O'Connor, B., & Smith, N. (2012). Censorship and deletion practices in Chinese social media. *First Monday*, *17*(3). https://doi.org/10.5210/fm.v17i3.3943

Bickert, M. (2020, January 6). *Enforcing against manipulated media*. Meta. Retrieved March 23, 2023, from https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/

Brooks, T., Princess, G., Heatley, J., Joseph, J. Kim, S., Parks, S., Reardon, M., Rohrbacher, H., Sahin, B., Spivak, James, S., Terrell., O., Richards, V. (n.d.). *Increasing Threat of DeepFake Identities.* U.S. Department of Homeland Security. https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf

Callon, M. (1984). Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay. *The Sociological Review, 34*(1), 196-233. doi/10.1111/j.1467-954X.1984.tb00113.x

Callon, M. (1987). Society in the making: The study of technology as a tool for sociological analysis. In W. Bijker, T. Hughes, & T. Pinch, (Eds.), *The Social Construction of Technological Systems.* (pp. 83-103). The MIT Press.

Cressman D. (2009). A brief overview of Actor-Network Theory: punctualization, heterogeneous engineering & translations. *Vancouver: ACT Lab/Center for Policy Research on Science & Technology School of Communication*, Simon Fraser University (Working Paper).

Denham, H. (2020, August 3). Another fake video of Pelosi goes viral on Facebook. *The Washington Post*. Retrieved from https://www.washingtonpost.com/technology/2020/08/03/nancy-pelosi-fake-video-facebook/.

Department of Infrastructure, Transport, Regional Development, Communications and the Arts. (2023, January 20). *New ACMA powers to combat harmful online misinformation and disinformation*. Ministers for the Department of Infrastructure. Retrieved March 23, 2023, from https://minister.infrastructure.gov.au/rowland/media-release/new-acma-powers-combat-harmful-online-misinformation-and-disinformation

Digi welcomes the government providing ACMA with oversight powers over misinformation. Digi. (2023, January 20). Retrieved March 23, 2023, from https://digi.org.au/digi-welcomes-the-government-providing-acma-with-oversight-powers-over-misinformation/

Educating Against Misinformation and Disinformation Act, H.R. 6971, 117th Cong. (2022). https://www.congress.gov/bill/117th-congress/house-bill/6971?s=1&r=46

Facebook. (2021). *Facebook response to the Australian disinformation and misinformation industry code*. https://digi.org.au/wp-content/uploads/2021/05/Facebook-commitments-under-disinfo-and-misinfo-code-final-report.pdf

Fallis, D. (2020). The Epistemic Threat of Deepfakes. *Philosophy & Technology*, *34*(4), 623-643. https://doi.org/10.1007/s13347-020-00419-2

Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2021). Deepfake Detection by Human Crowds, Machines, and Machine-informed Crowds. *Proceedings of the National Academy of Sciences*, *119*(2). https://doi.org/10.1073/pnas.2110013119

Hine, E., & Floridi, L. (2022). New deepfake regulations in China are a tool for social stability, but at what cost? *Nature Machine Intelligence*, *4*(7), 608–610. https://doi.org/10.1038/s42256-022-00513-4

Horton, J., & Sardarizadeh, S. (2022, July 28). False claims of 'deepfake' President Biden go viral. *BBC News*. Retrieved from https://www.bbc.com/news/62338593.

How meta prioritizes content for Review. Transparency Center. (2022, January 26). Retrieved March 23, 2023, from https://transparency.fb.com/policies/improving/prioritizing-content-review/

How technology detects violations. Transparency Center. (2022, January 19). Retrieved March 23, 2023, from https://transparency.fb.com/enforcement/detecting-violations/technology-detects-violations/

Hurcombe, E., & Meese, J. (2022). Australia's digi code: What can we learn from the EU experience? *Australian Journal of Political Science*, *57*(3), 297–307. https://doi.org/10.1080/10361146.2022.2122774

Lee, J., & Liu, C. (2016). Real-Name registration rules and the fading digital anonymity in china. *Washington International Law Journal*, *25*(1), 1–34.

Mervosh, S. (2019, May 24). Distorted Videos of Nancy Pelosi Spread on Facebook and Twitter, Helped by Trump. *The New York Times*. https://tinyurl.com/235jvaut

Meta. (2020, June). *Deepfake Detection Challenge results: An open initiative to advance AI*. Meta AI. Retrieved March 23, 2023, https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/

Meta. (2022). *Meta's Investments in Fact-Checking*. https://www.facebook.com/formedia/blog/third-party-fact-checking-industry-investments

Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., & Dwivedi, Y. K. (2023). Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research, 154*, 1–15. https://doi.org/10.1016/j.jbusres.2022.113368

O'Donnell, N. (2021). Have we no decency? section 230 and the liability of social media companies for deepfake videos. *University of Illinois Law Review, 2021(2),* 701-ii.

Pawelec M. (2022). Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions. *Digital society : ethics, socio-legal and governance of digital technology*, *1*(2). https://doi.org/10.1007/s44206-022-00010-6

Rana, S., Sung, A. (2020, August 1-3). *DeepfakeStack: A deep ensemble-based learning technique for deepfake detection* [Paper presentation]. 7th IEEE International Conference on Cyber Security and Cloud Computing, New York, NY, United States.

Rini, R. (2022). Deepfakes, Deep Harms. *Journal of ethics & social philosophy*, *22*(2). https://doi.org/10.26556/jesp.v22i2.1628

Smith, H., & Mansted, K. (2020). Weaponised deep fakes. *Weaponised deep fakes: National security and democracy* (pp. 11–14). Australian Strategic Policy Institute. http://www.jstor.org/stable/resrep25129.7

Twitter. (2020, February 4). *Building rules in public: Our approach to synthetic & manipulated media*. Twitter. Retrieved March 23, 2023, from https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media

Twitter. (n.d.). *Our synthetic and manipulated media policy | twitter help*. Twitter. Retrieved March 23, 2023, from https://help.twitter.com/en/rules-and-policies/manipulated-media

Williams, N. (2022). *Actor-Network Theory analysis of China's approach to regulating misinformation.* [Figure 4]. *STS Research Paper:* Deepfake's deep scars on democracy. (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Williams, N. (2022). *Actor-Network Theory analysis of the Australian approach to regulating misinformation.* [Figure 5]. *STS Research Paper:* Deepfake's deep scars on democracy. (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Williams, N. (2022). *Actor-Network Theory analysis of the proposed approach to regulating misinformation.* [Figure 6]. *STS Research Paper:* Deepfake's deep scars on democracy. (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Williams, N. (2022). *Actor-Network Theory analysis of the United States approach to regulating misinformation.* [Figure 3]. *STS Research Paper:* Deepfake's deep scars on democracy. (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Williams, N. (2022). *Handoff Model Diagram for Deepfake Technology.* [Figure 1]. *STS Research Paper:* Deepfake's deep scars on democracy. (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Williams, N. (2022). *Proposed Handoff Model for Deepfake Technology.* [Figure 7]. *STS Research Paper:* Deepfake's deep scars on democracy. (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA

Yang, F., Liu, Y., Yu, X., & Yang, M. (2012). Automatic detection of rumor on Sina Weibo. *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*. https://doi.org/10.1145/2350190.2350203

Zachary G. P. (2020). Digital Manipulation and the Future of Electoral Democracy in the U.S. *IEEE Transactions on Technology and Society, 1*(2). https://ieeexplore-ieee-org.proxy01.its.virginia.edu/document/9099201