DETECTING DEEPFAKED MEDIA WITH DEEP LEARNING

DEEPFAKES' DEEP SCARS ON DEMOCRACY

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Nathan Williams

October 27, 2022

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.

ADVISORS

Catherine D. Baritaud, Department of Engineering and Society

Briana B. Morrison, Department of Computer Science

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 standard of trustworthy media, one that was much harder to tamper with. Now, in the early 21st century, incredible progress in computer science and machine learning threatens the credibility of all forms of media. Deepfakes are a subset of synthetic media that use machine learning to create “believable, realistic videos, pictures, audio, and texts of events which never happened” (Brooks et al., n.d., p.3). The uses of deepfakes are almost endless and range from benign special effects in the film industry to democracy threatening fabrications on the internet. While some deepfakes that are made with minimal effort for blatant jokes can be easily detected by the human eye or ear, when given sufficient computing power and training data a surprising majority of people are fooled. Some studies even indicate that up to 85% of people can be fooled by a well-made deepfake (Dobber, 2020, p. 78). Combine this alarming statistic with the fact that there are already deepfakes of major politicians saying controversial statements, and there is a tangible threat to democracy that needs to be thoroughly considered.

The technical portion of this thesis, with Professor Briana Morrison as an advisor, will examine what aspects of a deepfaked video or audio provide the most reliable indicators of tampering, as well as investigate which classifier architecture(s) is most effective and efficient at classifying deepfakes. In a tightly coupled STS report, under the advice of Professor Catherine Baritaud, a handoff model and a technology and social relationships theory will be employed to investigate the consequences of deepfakes on political landscapes, first analyzing and classifying notable political deepfakes based on content, target, and message, then evaluating responses to malicious deepfakes from legislators, corporations, and individual politicians. The project is set to be completed in the spring semester of 2022 as shown below in Figure 1.

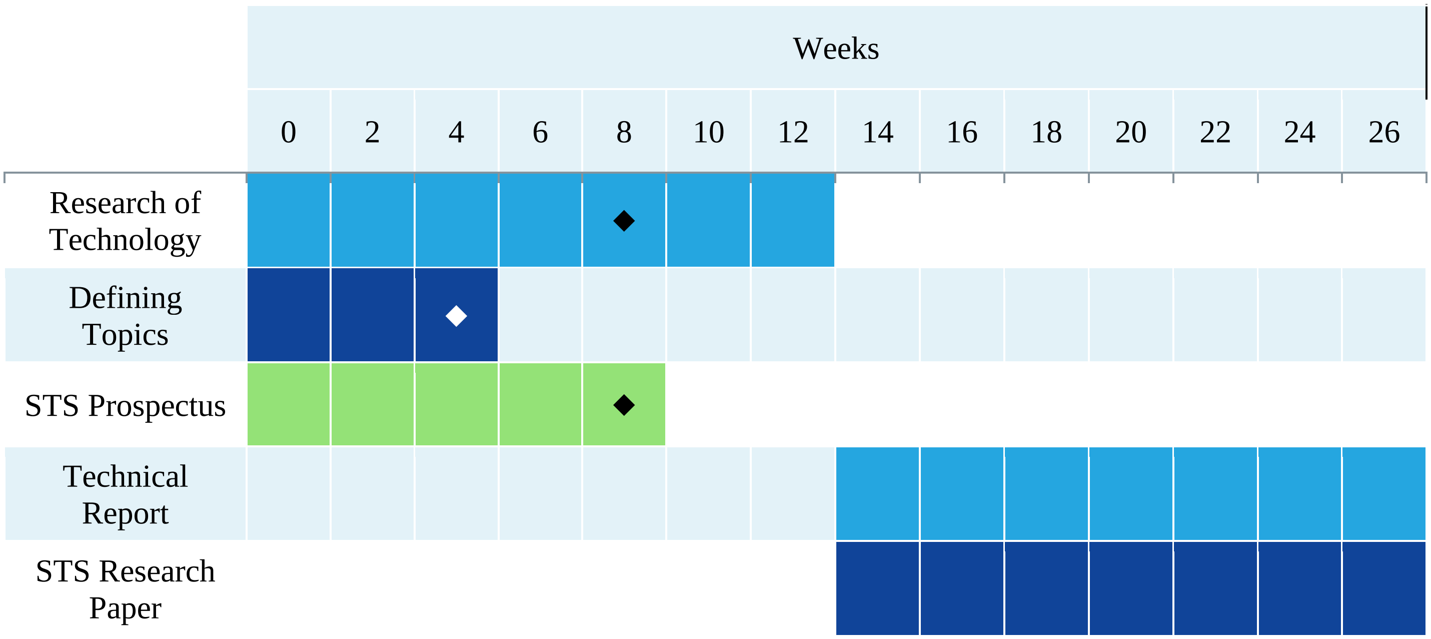

Figure 1: Gantt Chart for Computer Science Capstone. This figure demonstrates the expected timeline for the capstone deliverables over two 14-week semesters. (Williams, 2022).

The technical report will be completed in the CS 4991 class, and the STS paper will be written throughout the STS 4600 class.

DETECTING DEEPFAKED MEDIA WITH DEEP LEARNING

The entire purpose of deepfake technology is to be undetectable; to be so realistic that the viewer will not question its legitimacy. What makes deepfakes even harder to detect is that once the video is rendered, there are no flags in its file data that would immediately 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). As such software for detecting deepfakes needs to analyze the video as presented rather than checking the file data. One of the leading methods in the field of detecting deepfakes is using deepfake detection classifiers (DDC), or deep learning algorithms that separate given data into categories: deepfake or legitimate (Rana et al., 2022, p. 25508). The accuracy of a DDC is based on many factors, but two of the most significant are the architecture of the classifier and the artifacts used (Rana et al., 2022, p. 25508). The architecture determines the structure of the neural network and which algorithms are used for data acquisition, processing, modeling, and eventual execution of the classifier. The artifacts are what the classifier is "looking for" in the given video to use as reference for its decision on if the video is fake or not (Rana et al., 2022, p. 25506). Since the architecture determines how the artifacts are mapped and evaluated, these two aspects of a deepfake detection go hand in hand and anyone developing a deepfake detection classifier cannot implement one without considering the other.

COMPLICATIONS MOTIVATING FURTHER RESEARCH

While the standard approach to deepfake detection involves choosing a classifier architecture suitable for the data set and selecting artifact(s), recently, researchers have proposed using a combination of architectures to improve the accuracy the model. Figure 2 below shows how David Guera and Edward Delp (2018), distinguished professors of computer engineering at Purdue University, used a convolutional neural network (CNN) to create an artifact based on facial features (p. 4). Guera and Delp (2018) then expanded on the standard DDC model by also implementing another deep learning architecture known as a long short-term memory (LSTM) neural network to create a new artifact that uses the feature vectors from the CNN to measure temporal inconsistencies between frames rather than measuring facial features (p. 4). This classifier achieved a significantly higher accuracy rate for detecting deepfakes than the standard single architecture approach (p. 4). However, not all deep learning models and architectures are compatible, and so the introduction of training based on multiple artifacts and creating artifacts with other deep learning models opens new avenues for research into the combination of different deep learning models and specialized artifacts to create a more accurate classifier (Rana et al., 2022, p. 25508).

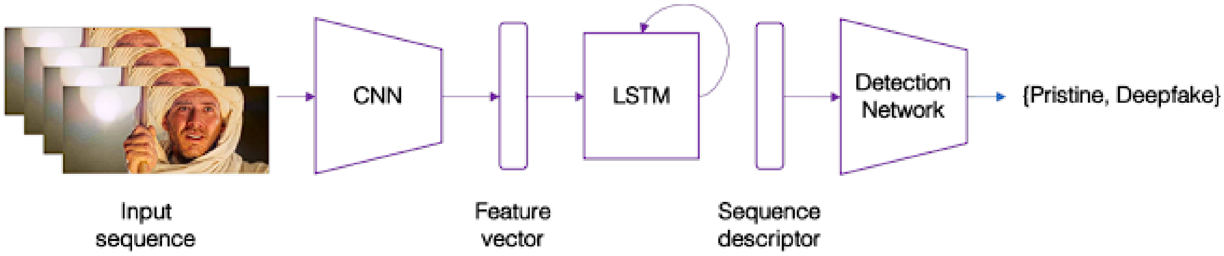

Figure 2: Combined Architecture Deepfake Detection Classifier. A diagram of a deepfake detection classifier, showing how each architecture is used to develop a detection network. (Guera & Delp, 2018, p. 4).

MODES OF RESEARCH AND ANTICIPATED OUTCOMES

The documents that will be consulted when researching the technical portion of this project are academic and scholarly articles as well as relevant open-source code. These resources will be referenced in a state-of-the-art technical report that will extend the field of artificial intelligence and deep learning in computer science. The goal of this report is to investigate how the type and number of artifacts used in a deepfake detection classifier affects the accuracy of a deepfake detection classifier and offer recommendations for artifacts in the context of notable classifier architectures.

DEEPFAKES' DEEP SCARS ON DEMOCRACY

The STS research topic is tightly coupled with the detection of deepfakes and focuses on how deepfakes are used politically as well as how society is responding to this new form of 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 a deepfake of Nancy Pelosi, speaker of the United States House of Representatives, portrayed her giving an interview while slurring her words and appearing drunk. 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 an enormous amount of his followers. While many other political deepfakes aim to slander a political opponent, some are used with the intent of falsely bolstering a politician's reputation, such as one deepfake which was edited to show Pope Francis endorsing Donald Trump for president (Allcot & Gentzkow, 2017). With political deepfakes falsely portraying politicians in both positive and negative lights, they represent a tangible threat to society, and the first step in addressing this misinformation is to understand how it is used. Recognizing the content, target, and message of a politically affiliated deepfake is essential to discovering patterns in malicious deepfakes that can be used as a basis for creating societal safeguards against misinformation.

Developing social safeguards against misinformation is critical to democracy because while researchers continue to make strides in deepfake detection technology, deepfakes are still being circulated online. Through the Communications Decency Act of 1996 (CDA), the United States government protected "online service providers from legal liability stemming from content created by the users of their services" (Zachary, 2020, p. 109). This act was passed before deepfakes were created, but the law still stands today, removing responsibility from service providers of monitoring their own platforms for deepfaked media and other misinformation. Rather than revisit the CDA, legislators have preferred to allow social pressure to convince some companies like Facebook and Instagram to implement third-party fact checkers

(Instagram, 2019). However, these third-parties fact checkers cannot reach all the posts flagged for misinformation in a timely manner and without detailed legal consequences many companies are slow to remove altered media. Xiaoli Nan (2022), a professor in the Department of Communications at the University of Maryland, emphasizes how even after misinformation is removed or marked as false it still has, and still is influencing consumers (p. 2). As such, further research, and discussion on legislation is essential to improving and hastening the process of detecting deepfaked media.

THE EXTENT OF DEEPFAKES' HARM

On the analysis of the content, target, and message Regina Rini (2022), a professor and research chair in Philosophy of Moral and Social Cognition at the University of Wisconsin offers a complicating factor to the political use of deepfakes with the notion of individual harm. While it may seem obvious, when considering political deepfakes at a societal level it is easy to forget that they can be used for more than just spreading misinformation. "Frakenporn" or pornography generated from deepfakes can be used for blackmail on political candidates, and panoptic gaslighting could be used to convince a politician that they took a stance or spoke on an issue they never had before (p. 143-7).

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). Pawlec (2022) even mentions an example where speculation that a video of

Gabonese President Ali Bongo was fabricated contributed to a failed military coup (p. 12). The idea of reversing the intent of deepfakes and proclaiming real events as fabrications rather than making fake events seem real is a troubling offshoot of this technology and threatens societies shared reality. A shared reality refers to the idea that all members of a society accept certain facts and series of events as true (Zachary, 2020, p. 110). Without a shared reality, proper discussion, debate, and compromise becomes virtually impossible and severely limits the capabilities of a democracy, which only emphasizes the need for proper legislation and protocols surrounding deepfakes (Zachary, 2020, p. 110).

Furthermore, in terms of governmental response to deepfaked media, the Educating Against Misinformation and Disinformation Act (EAMDA) was recently introduced to Congress, which proposes citizen education on misinformation and deepfaked media as a partial solution to the spread of fabricated media (Educating Against Misinformation and Disinformation Act, 2022). While legislation offers a promising angle to combat political deepfakes, none of the proposed education plans pitched in the EAMDA were mandatory and it is unclear how far reaching the bill will be if passed (Educating Against Misinformation and Disinformation Act, 2022). With legislation around deepfakes and other misinformation noticeably lacking enforcement capabilities, analyzing the strategies and effectiveness of corporate and governmental responses to misinformation could offer insight into improving legislation and standards.

MISINFORMATIVE ENVIRONMENTS

The STS research paper will be written with sources from public social media records, official legislation documents, academic articles, and publicly available deepfakes. On top of investigating current uses of political deepfakes and practices for mitigating misinformation damage, this paper will outline potential future uses of political deepfakes as the technology is improved. The STS Handoff framework in Figure 3 will be applied to show how deepfake technology is transferred from software engineers to malignant actors, and eventually to the public or to politicians. This Handoff framework will help to identify "weak points" in the journey of a deepfake where legislation or corporate practices can have a meaningful impact on mitigating the damages done by a malicious deepfake.

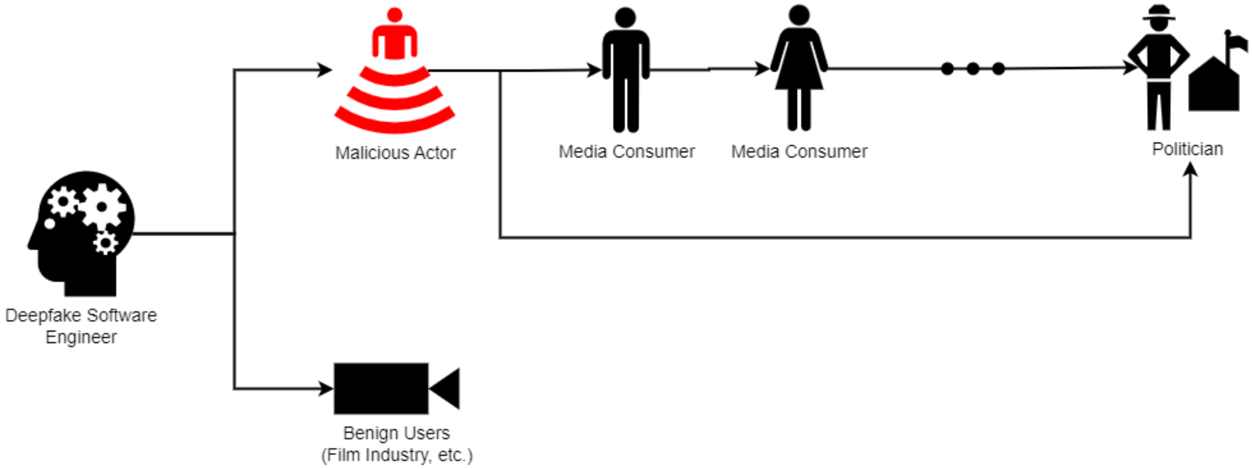

Figure 3: Handoff Model for Deepfake Technology. This figure demonstrates the journey of deepfake technology as it spreads throughout society. Adapted from STS Frameworks, by Carlson, W. B., 2009.

Figure 3 shows how there could be any number of media consumers between the initial release of the deepfake and the targeted politician, meaning there could be any number of people influenced by the deepfake. As the deepfake is shared from consumer to consumer, it is repeatedly saved and compressed, thus lowering the overall quality of the video, and making it even more difficult to pick up on the minute details that give away a deepfake. Furthermore, malicious actors can directly hand off the deepfake to a politician in the case of blackmail or panoptic gaslighting.

A supplementary STS model that can be implemented for this topic is the Technology and Social Relationships (TSR) framework shown in Figure 4. A TSR framework is incredibly helpful in visualizing the intents and content of a deepfake from a malicious actor by analyzing the relationships with the intended targets of the deepfake as well as the engineer who provides the technology.

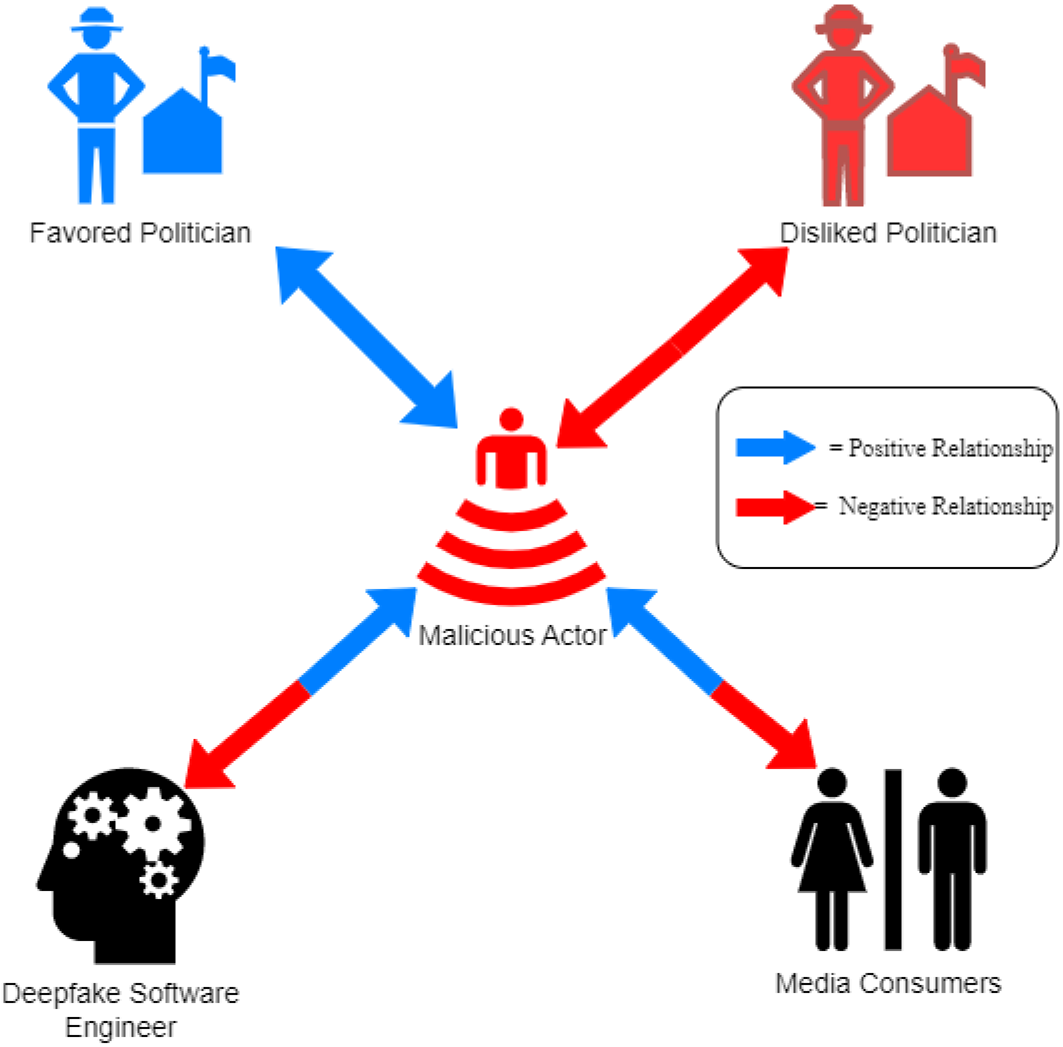

Figure 4: Technology and Social Relationships Diagram for Deepfake Technology. This figure demonstrates the relationships of a malicious actor with their intended targets and software engineer. Adapted from STS Frameworks, by Carlson, W. B., 2009.

Figure 4 shows how a malicious actor creating a political deepfake can either slander a disliked politician or bolster the reputation a favored politician. In return, a politician negatively affected by a deepfake may push for legislation limiting their capabilities, but a politician positively affected may dismiss political deepfakes as a threat. Furthermore, by sharing the deepfake or even just viewing it, media consumers help the malicious actor by spreading their deepfake while the malicious actor attacks their shared reality. Finally, A deepfake software engineer provides the technology necessary for a malicious actor to create a political deepfake, and the malicious actor perverts the engineer’s work.

KEEPING REALITY REAL

Deepfakes are a testament to the heights of computer science, the ability to realistically fabricate media is something that software engineers should be proud of from a technical standpoint. However, the social consequences of deepfakes have the potential to devastate any and every society. When reality erodes into uncertainty, when video evidence and spoken word can no longer be trusted, confusion will run rampant and dialogue between political parties and even individual people will become increasingly difficult without a shared reality to ground a conversation in. As such, it is imperative that governments, corporations, and engineers combine efforts to research both social and technological strategies to address this growing threat. Using the same technology that drives deepfakes could help to classify them with reasonable certainty and enforcing strict but ethical regulations on altered media could help society regain and maintain a shared reality.

REFERENCES

Allcott, H., Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, *31(*2), 211-36. https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211

Baritaud, C. & Carlson, W. B. (2009). STS Frameworks. [Figure 3 & Figure 4]. Class handout (Unpublished). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

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

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2020). Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics*, 26(1), 78. https://journals.sagepub.com/doi/full/10.1177/1940161220944364

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

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

Guera, D., Delp, E. (2018). *Deepfake video detection using recurrent neural networks* [Paper presentation]. 15th IEEE International Conference on Advanced Video and Signal Based Surveillance, Auckland, New Zealand.

Instagram. (2019, December). *Combatting Misinformation on Instagram*. Instagram. https://about.instagram.com/blog/announcements/combatting-misinformation-on-instagram

Mervosh, S. (2019, May 24). Distorted Videos of Nancy Pelosi Spread on Facebook and Twitter, Helped by Trump. *The New York Times*. https://www.nytimes.com/2019/05/24/us/politics/pelosi-doctored-video.html

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., Nobi, M., Murali, B., & Sung, A. (2022). Deepfake Detection: A Systematic Literature Review. *IEEE Access*, *10*. https://ieeexplore.ieee.org/document/9721302/authors#authors

Rana, S., Sung, A. (2020, August 1-3). *DeepfakeStack: A deep ensemble-based learning technique for deepfake detection* [Paper presentation]. 7th IEEE Internation 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

Williams, N. (2022). *Gannt Chart for Computer Science Capstone.* [Figure 1]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Williams, N. (2022). *Handoff Model Diagram for Deepfake Technology.* [Figure 3]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Williams, N. (2022). *Technology and Social Relationships Diagram for Deepfake Technology.* [Figure 4]. *Prospectus* (Unpublished undergraduate thesis). School of Engineering and Applied Science, University of Virginia. Charlottesville, VA.

Nan, X., Wang, Y., & Thier, K. Why do people believe health misinformation and who is at risk? A systematic review of individual differences in susceptibility to health misinformation. *Social Science & Medicine, 314.* https://www-sciencedirect-com.proxy01.its.virginia.edu/science/article/pii/S0277953622007043#!

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