An Enhanced Machine Learning Model to Detect Deepfakes

An Analysis of Legislative Measures to Combat Deepfake Pornography

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
Timothy Cha

May 3, 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

Kathryn A. Neeley, Department of Engineering and Society

Introduction (250-350 words)

When eighteen-year-old Noelle Martin did a reverse Google image of herself, she discovered her face superimposed on the bodies of several pornography stars – though she was never involved in such acts (Gieseke, 2020). Martin's case is an example of the horrifying phenomenon of deepfake pornography, increasingly rampant due to the rapid advancements of technology. Men all over the Internet have damaged the sexual privacy of women, and deepfake pornography has grown to become a new outlet of such violation (Gieseke, 2020). Deepfakes originated when a Redditor under the pseudonym "u/deepfakes" used Google-given machine learning software to superimpose celebrities' faces on the bodies of pornography stars, initiating a movement where users could distribute deepfake pornographic material of people who they know well onto online forums like Reddit and Discord (Harris, 2019). This problem is critical to study because given that deepfakes are continually evolving to blend in with organic material, society must become more aware of the implications of deepfakes and those involved in the distribution of deepfake material must be held accountable. Research towards deepfake detection, in addition to legislative measures which can hold accountable the different agents of deepfake distribution, can contribute towards reducing the number of victimizations instigated by deepfake pornography and ultimately destroying deepfakes in general. My technical research paper will focus on analyzing improved algorithms for detecting deepfake material and developing a unique algorithm to detect deepfakes, while my STS research paper will investigate current legislative measures of deepfake pornography and determine which legal remedies may be the most effective.

Technical Topic: An Enhanced Machine Learning Model to Detect Deepfakes (400-500 words)

My technical topic will explore the development of machine learning algorithms which detect deepfakes and will propose an alternative deepfake detection model in contrast to more traditional ones. This problem is important to study because current machine learning models which perform deepfake detection have shortcomings (Jung, 2020), and given how indistinguishable current deepfakes can be from original videos, deepfake attacks like pornographic material can be interpreted as organic and hence can have serious implications on the lives of victims. An enhanced deepfake detection model which can endure continually evolving deepfakes can be a powerful tool in holding deepfake creators accountable.

Current research is being done on deepfake detection models. Generative Adversarial Networks are a type of machine learning model which create and detect deepfakes. (Jung, 2020) states that deepfakes work by "iterating an actual data-based generation and verification task through two opposite deep learning models." However, (Jung, 2020) also notes that a popular method of detecting deepfakes, detection of the collapse of pixels, has become fallible due to more recent deepfakes which can overcome such limitations. (Jung, 2020) and his team have proposed to determine whether a certain material is a deepfake or not based on the eye blinking of the subject in a deepfake. Despite the lack of pattern in eye blinking, I am concerned that deepfakes will eventually be able to synchronize the blinking of the human eye with the subjects of the deepfake. Thus, instead of relying on only one frame to differentiate between original and deepfake material, (Amerini et. al, 2019) have proposed an alternative method of determining whether a source of material is original or deepfake using *optical flow fields*, and (Amerini et. al, 2019) attempt to identify potential flaws in between frames of a deepfake video.

My research aims to follow a similar method, where I attempt to identify a rather unique way of detecting deepfake material. Like other researchers, I will use Generative Adversarial Networks to create my own deepfake detection model, but pinpoint other features besides eye movement and inter-frame changes that are solely unique to deepfakes. I hope that through my research I can create a model which can withstand the constant evolution of deepfake material, and potentially identify any distinguishing factors between the original source material and the deepfake. I plan to use non-pornographic deepfake material as data to feed into my machine learning model, which I then will gather data on the model's accuracy in differentiating between deepfakes and organic material. To assist my research, I intend to reach out to UVA professors who specialize in Machine Learning, specifically professors Rich Nguyen and Jane Qi.

STS Topic: An Analysis of Legislative Measures to Combat Deepfake Pornography (500-750 words)

My STS topic analyzes current legislative measures to combat deepfake pornography and proposes what I think is the best legal remedy to this phenomenon. Because of the recent emergence of deepfake pornography, legislative regulations for deepfake porn have not been fully developed (Harris, 2019). It is imperative that this issue is addressed sooner so that clear legal regulations are established which can contribute to an overall decrease of deepfake pornography related victimizations. Throughout my STS paper, I intend to review current legislative measures to this issue and determine whether stricter regulations to significantly reduce, if not eliminate, the prevalence of deepfake pornography are necessary. I choose to frame this issue in the Actor-Network Theory framework because I can view how each actor/actant of deepfake pornography contributes to the overall issue and which actors will need to be held most liable when it comes to introducing new legislative measures. Other frameworks such as Social Construction of Technology or Ethics of Care would not be effective when attempting to view the issue from a legal perspective because the relationships between the different actors of deepfakes are crucial to determining which demographics be held most accountable when creating new legislation.

Current scholars who have researched this problem have proposed varying solutions on the optimal legal remedy to deepfake pornography. (Harris, 2019) and (Gieseke, 2020) both agree that a federal statue is necessary to best remedy deepfake pornography. (Gieseke, 2020) further elaborates that platforms which allow such material should be held legally liable in addition to the perpetrator (Gieseke, 2020), considering that everyday Open Source Software platforms like GitHub passively contain repositories of deepfake pornography code (Newton and Stanfill, 2019). However, both (Harris, 2019) and (Gieseke, 2020) acknowledge that a potential federal statue may not be entirely effective because of First Amendment rights. On the other hand, scholars such as (Karasavva and Noorbhai, 2021) argue that current laws can be sufficient in remedying deepfake pornography and do not mention instituting a strict federal regulation which bans deepfake pornography entirely. Some authors, such as (Gibson, 2020), take an intermediate position and state that while current laws may have potential, a statutory solution may be better, though deciding which solution to go with may ultimately be up to time. Through my research, my goal is to determine which of these two opposite ends of this spectrum will have a greater effect on regulating deepfake pornography in the longer term, and how the conflicting issues of First Amendment rights and the violation of sexual privacy can be mitigated.

For my STS research, I will collect evidence primarily from court cases with issues similar to deepfake pornography, as there have not been any document court cases where the victim accused the perpetrator of distributing a deepfake video of themselves. Case studies are best because they provide a detailed view of the incident, what goes on inside the mind of the victim, and how other actors respond to the incident. To add to my insight on this issue, I also intend to interview with experts of deepfakes as well as lawyers who may deal with such cases.

Conclusion (150 words)

Throughout this research process, I hope to significantly contribute towards the development of deepfake detection neural networks with an enhanced machine learning model that can accurately identify deepfake material and can withstand technological advancements in deepfakes. I hope that my machine learning model receives widespread support from researchers in the same field and that it can eventually become tested on deepfake videos frequently. Eventually, I hope that my model achieves practical uses in court cases to defend victims of deepfake pornography. I also hope that through my STS paper that I can propose what I think is an effective legal strategy to uphold different agents of deepfake pornography accountable, which can be heard by legislators and incite them to start taking action to remedy deepfakes. Most importantly, I hope that my research contributes towards the ultimate cease of deepfake pornography as a phenomenon.

References

Amerini, I., Galteri, L., Caldelli, R., & Del Bimbo, A. (2019). Deepfake Video Detection through Optical Flow Based CNN. 0–0. https://openaccess.thecvf.com/content_ICCVW_2019/html/HBU/Amerini_Deepfake_Video_Detection_through_Optical_Flow_Based_CNN_ICCVW_2019_paper.html?ref=https://githubhelp.com

Burkell, J., & Gosse, C. (December 2, 2019). Nothing new here: Emphasizing the social and cultural context of deepfakes. First Monday. https://doi.org/10.5210/fm.v24i12.10287

Gibson, K. (2020). Deepfakes and Involuntary Pornography: Can Our Current Legal Framework Address This Technology? Notes. Wayne Law Review, 66(1), 259–290.

Gieseke, A. (2020). "The New Weapon of Choice": Law's Current Inability to Properly Address Deepfake Pornography. Vanderbilt Law Review, 73(5), 1479-1516.

Harris, D. (2018-2019). *Deepfakes: False Pornography Is Here and the Law Cannot Protect You.* Duke Law & Technology Review, 17, 99-128.

Jung, T., Kim, S., & Kim, K. (2020). DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern. IEEE Access, 8, 83144–83154. https://doi.org/10.1109/ACCESS.2020.2988660

Karasavva, V., & Noorbhai, A. (2021). The real threat of deepfake pornography: a review of Canadian policy. Cyberpsychology, Behavior, and Social Networking, 24(3), 203-209. https://doi.org/10.1089/cyber.2020.0272

Maddocks, S. (2020). 'A Deepfake Porn Plot Intended to Silence Me': Exploring continuities between pornographic and 'political' deep fakes. Porn Studies, 7(4), 415–423. https://doi.org/10.1080/23268743.2020.1757499

Newton, Olivia B., and Mel Stanfill. "My NSFW Video Has Partial Occlusion: Deepfakes and the Technological Production of Non-Consensual Pornography." Porn Studies 7, no. 4 (December 09, 2019): 398–414. https://doi.org/10.1080/23268743.2019.1675091.

Öhman, C. (November 19, 2019). Introducing the pervert's dilemma: A contribution to the critique of Deepfake Pornography. Ethics and Information Technology, 22(2), 133–140. https://doi.org/10.1007/s10676-019-09522-1