Thesis Portfolio

The Role of Software Engineering in Providing Funding for After-School Programs

(Technical Report)

Deepfakes and Social Media

(STS Research Paper)

An Undergraduate Thesis

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

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

Author
Siddharth Ghati
May 08, 2020

Table of Contents

SocioTechnical Synthesis

The threat that Deepfakes, realistic videos and pictures created by machine learning algorithms, pose to information consumers requires a change in current content moderation policies on social media platforms. Currently much of the content moderation on these platforms is outsourced to human moderators. However, this structure becomes an apparent flaw when dealing with Deepfakes due to the fact that Deepfake artifacts are designed to fool humans. As a result, automated solutions such as deep-learning based detection tools can be used as possible countermeasures against Deepfakes. Using these technologies would delegate the responsibility of content moderation from humans to automated technologies. As a result, the primary STS theory that I will apply is Latour's Actor Network Theory (ANT). I will be specifically focusing on the delegation of responsibility to technology.

As with all discussions of automation, the human and social factors need to be considered and this is no different. The biggest impact that this move to automated moderation would have is the loss of jobs for many human moderators in foreign countries. In addition, the creators of the detection technologies and what they do with their technology are important. This is due to the fact that any detection algorithm would essentially be playing a game of cat and mouse with any creation algorithm as any technique that is used to detect Deepfakes can also be used to create better Deepfakes. So, for example, if the creators of the detection technologies decide to leave their technology open source then Deepfakes will only get past the detection algorithm's techniques. As a result, I will use ANT again to analyze the relationships between the detection technology, the creators of the technology, malicious agents manufacturing Deepfakes and human moderators.

In order to conduct my STS research, I will be using Google Trends to understand just how many people understand what Deepfakes are. I also plan on studying cases of fake information spreading on the internet through social media to find patterns across various cases. In addition, I have already started looking at current content policies on social media platforms and current research being done to detect Deepfake artifacts. Through my research I hope to understand how well informed my local population is about Deepfakes and find potential solutions that social media platforms can implement to combat against Deepfakes. Looking at my STS research and the technologies that I research together a clearer picture of how social media can be used to spread misinformation and some solutions to stop this appears. By better understanding misinformation and what can be done to stop it we can work towards a better future where you can believe anything you read on the internet.