Prospectus

After School Association of America Web Application
(Technical Topic)

Deepfakes and Social Media Content Management Policies
(STS Topic)

By

Siddharth Ghatti

10/30/19

Technical Project Team Members: Jack Durning, Nadia Hassan, Tae Whoan Lim, Victor Cruz

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.

Signed: Siddharth Ghatti

Approved: _____ Date _____
Rider Foley, Department of Engineering and Society

Approved: Ahmed Ibrahim Date 11/25/2019
Ahmed Ibrahim, Department of Computer Science

Rider Foley
Technical Topic
Description of After-School Association of America

The After-School Association of America is a nonpartisan association created to assist schools who provide services to students' after-school through various programs. Their mission is to provide financial assistance to schools in an effort to increase participation of students after normal school hours. They exist to remove financial barriers in the way and transportation issues for students and their parents so more students can participate in and enjoy school activities. "We are a national organization working with schools to prevent gang affiliation by increasing the number of youths supported during out-of-school time".

A major problem they have faced is a lack of infrastructure. The non-profit is run by the founder Michelle Busby with little to no help. She personally gathers and refines the data she requires from each individual she works with. The company processes applications requesting financial aid for after-school activities. The funding can be used for transportation, staffing, and activity materials and many more. The non-profit receives the capital required to fulfill these application requests from individual donations and grants that the founder personally applies for. All of this data is also handled by the founder alone. This current solution has a single person trying to manage all of this data. This data is currently collected over the phone and through email. These elicitation methods require a lot of time and energy guiding the user through the process. This data is then stored on a personal computer in a variety of formats. The current system for the organization has no structured way of allowing those who seek funding, the primary users, nor organization administrators to register an account and manage their side of the funding process directly through an interface. Given that there is no way for the users who provide funding donations to register into a user database to continue to provide funds, participating schools cannot submit detailed applications and information for funding sources offered by ASAA. Furthermore, leaders of ASAA cannot manage user accounts created nor process applications with all the necessary materials needed to affirm funding eligibility in an effective and timely manner.

Why Our System is Important

Our application system will aggregate all of this data into a single access point that the founder can use to better collect, track, and utilize this data. The incorporation of multiple types of users becomes useful for all parties involved when it comes to ASAA and the schools it works with since all users can manage programs and funding through the interface itself. Our system also guides users through the entire application process by prompting users for all the necessary information and ensuring that the information is sent to the key administrators of the organization. In this way, our system will streamline the entire application process for the user as well as the owner. Overall, our system will be more efficient, safer, and easier. Ultimately, these innovations will encourage more schools and student participants to reach out for assistance in funding to ensure that every school has the opportunity to build successful afterschool programs.

What our System Does

Our system allows users to create applications so that owners and supervisors can approve applications, deny them, or request for more information. Our system puts the entire process in one easy-to-use location. The system will have the ability to have several different user classes for the purposes of providing funds, receiving funds through an application creation portal that allows users (schools and students) to create applications in a simplified manner, and for administrators to approve, deny, or request further action when reviewing applications and managing funds allocated to users who receive funding. The system will integrate a central database with encrypted information for security purposes to easily monitor school and funding data for all parties who use the new user interface. Lastly, our system will also analyze the data inputted into the database and provide useful statistics for the founder. These statistics will definitively show the effectiveness of this company.

System Requirements

Gathering system requirements is important because it is how you find out what it is you need to build. If you are not careful and thorough in your requirements elicitation, you will end up building the wrong product. During this process, developers are meeting with a client who may not be familiar with technology. As a result, one must be very careful to pay attention and talk clearly to the client. Even when requirements elicitation is done correctly, the build might still not be what the customer wanted. Therefore, one must keep an open and direct communication channel with the client at all times.

Minimum Requirements:

As a USER:

I should be able to subscribe by providing (“Email”, “First and Last Name”, “Phone Number”)

I should be able to create a new application

I should be able to upload documents to the application.

I should be able to check the application status.

I should be able to receive application status updates.

I should be able to notice that I have subscribed.

I should be able to opt out of a subscription.

I should be able to notice that the application has started with a link to return to complete the application.

I should be able to notice that the application was received.

I should be able to notice that more documents are needed.

I should be able to receive a notice of an approval.

I should be able to receive a notice of a denial.

I should be able to return to an application.

I should be able to cancel an application.

I should be able to edit an application after submission.

I should be able to edit the application up to a week prior to the deadline. After that any changes can only be performed by an admin.

I should be able to send an email to the admins.

I should be able to refer other users.

I should be able to leave comments.
I should be able to leave a review.

Desired Requirements:

As an ADMIN:
I should be able to review applications.
I should be able to cancel applications.
I should be able to create a new application on a user's behalf.
I should be able to edit a user application.
I should be able to email applicants.
I should be able to manually add users.
I should be able to remove users.
I should be able to grant user access manually.
I should be able to send requests to users for additional information.

As a SUPER USER:
I should be able to review applications.
I should be able to cancel applications.
I should be able to email applicants.
I should be able to manually add users.
I should be able to remove users.
I should be able to grant user access manually.
I should be able to send requests to users for additional information.
I should be able to add admins.
I should be able to deny applications.
I should be able to move application to the next step.
I should be able to approve an application.
I should be able to deny an application.
I should be able to export data (similar to owner).

As an OWNER:
I should be able to review applications.
I should be able to cancel applications.
I should be able to email applicants.
I should be able to manually add users.
I should be able to remove users.
I should be able to grant user access manually.
I should be able to send requests to users for additional information.
I should be able to add admins.
I should be able to deny applications.
I should be able to move application to the next step.
I should be able to approve an application.
I should be able to deny an application.
I should be able to pull data.
I should be able to add funds received by the source with the date.

Optional Requirements:

As an OWNER:

I should be able to make charts, graphs, and spreadsheets with data.

STS Topic

Just as our technical project improves upon previous generations of the same system, so do Deepfakes. Specifically, Deepfakes that are generated by Generative Adversarial Networks (GAN) improve upon from the mistakes of their past iterations. 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).

The realism of the Deepfake artifacts created by GAN's creates an issue for social media platforms aiming to eliminate the spread of artificial information 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). From interviewing some of the moderators, Dwoskin (2019) found that the nature of these jobs leaves the moderators traumatized without any mental-health resources to recuperate. The current model of content moderation has humans, at a large mental cost, 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

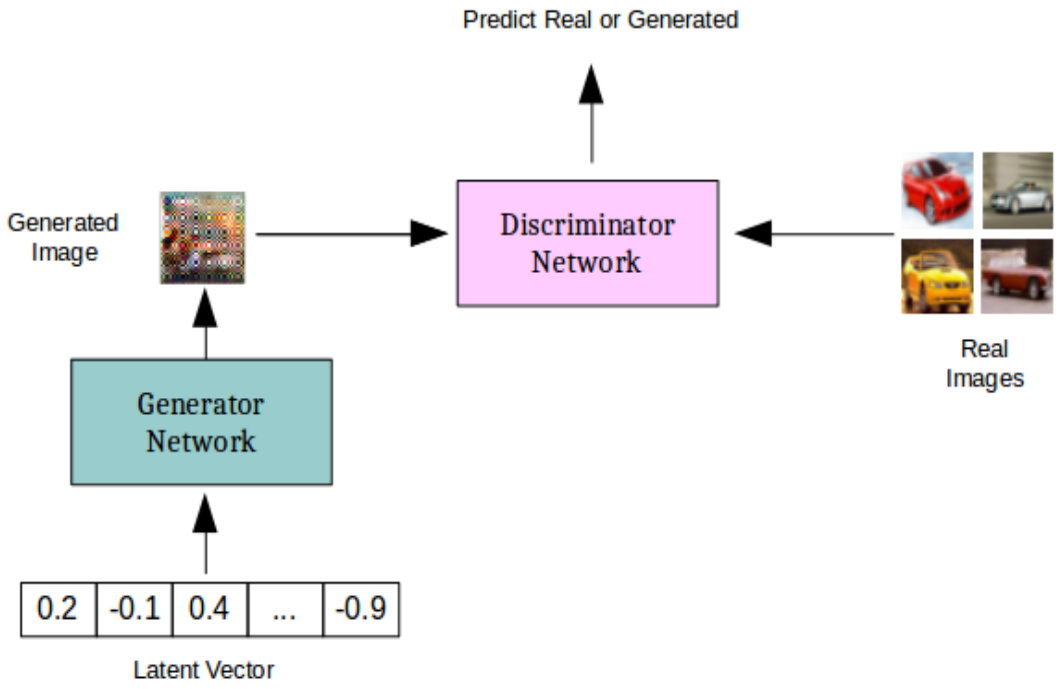

Figure 4. Generative Adversarial Network (Desai, 2018)

the need for social media platforms to deploy technological solutions to deal with Deepfakes. Going a step further, a case can be made for the delegation of 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).

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 people 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 behaviour 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).

One possible avenue for this delegation of content management to non-human agents is found with research done by David Guerra and Edward Delp (2018). In their research, Guerra and Delp designed a system that uses 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 assigning 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.

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 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. In fact, picture of the "Napalm Girl" was used in training for moderators as a picture that must be removed from the platform (Gillespie,2018). Although the image of "Napalm Girl" is not fake content, the same criticisms of censorship could arise against social media platforms if these platforms take a stringent stance when dealing with the removal of Deepfakes. On the other hand, if social media platforms take a

very lenient approach to removing Deepfakes and take no action to prepare for the spread of such content, then the bleak future where information can be manufactured by nefarious-agents to achieve political agendas becomes a very real possibility.

Research Question and Methods

With this, it can be seen that Deepfakes will drive social media platforms to make decisions about the very nature of their policies. Deepfakes will drive social media platforms to decide whether they want to be aggressive, lenient, or somewhere in the middle. This leads to the main question that will drive my research: How will Deepfakes impact social media platform's content management policies? Since one of the major actors in the question are the 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 social media platforms such as Twitter, Facebook, and Reddit. For example, Reddit's content policy does not allow for content that impersonates someone in a misleading or deceptive manner ("Reddit Content Policy", n.d). This indicates a formal stance against fake information on the platform, in whatever form it comes in. In addition to this, in order to gauge the current state of Deepfake technology, I will run a survey in which I will ask UVA students and staff to differentiate between Deepfake videos and original videos. Finally, I will 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, such as the system that was designed by Guerra and Delp (2018).

In addition to this, I will be looking case studies of fake stories that gained traction on social media platforms. One such example of this is the influx of fake information that gained traction during the 2016 presidential election. Hunt Allcoat and Matthew Gentzkow (2017)

found that 115 pro-Trump fake stories were shared a total of 30 million times and 41 pro-Clinton fake stories were shared a total of 7.6 million times on Facebook. By looking at these case studies, I aim to find similarities that exist between the instances of misinformation explosions and ideally be able to find solutions that social media platforms can implement to fight them. All of this being will be done with the underlying assumption being that Deepfake based misinformation will also follow similar patterns of spreading that current artifacts of misinformation do.

Timeline and Expected Outcomes

The expected outcome for our technical project is to have a working web application in production that will be tailored to the specifications and requirements that were given to us by the client at the beginning of this semester. The write up for this project will document the development process of the application. As for my STS topic, the expected outcome is to have a thoroughly detailed research paper that will address both the technological and social aspects of the topic. This will be done by drawing from the knowledge bases of both STS and Computer Science in order to flesh out all possible ways that Deepfakes can impact social media content management policies. The paper will also detail any possible solutions for the social media platforms that would allow the platforms to combat against malicious Deepfake. I also expect to gather primary evidence relating to the current state of Deepfake technology.

Task	**Expected duration of task**
Policy analysis of content management policies	01/22/20 – 02/06/20
Conduct Survey	2/13/20 – 2/22/20
Write STS research paper	01/22/20 – 04/01/20

Conduct research for the case studies of the spread of fake information as well as papers in the field of Deepfake technology.	01/22/20 – 04/01/20

Table 1. Timetable for STS research paper (Created by: Siddharth Ghatti, 2019)

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

Brownlee, J. (2019, August 14). A Gentle Introduction to Long Short-Term Memory Networks by the Experts. Retrieved from https://machinelearningmastery.com/gentle-introduction-l ong-short-term-memory-networks-experts/.

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

Django. (2020). Retrieved from https://www.djangoproject.com/.

Elizabeth Dwoskin, 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/.

Gillespie, T. (2018). Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press. pp1-5.

Guera, D. J., & Delp, E. J. (2018). Deepfake Video Detection Using Recurrent Neural Networks. 2018 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.

Nikolov, A. (2017, April 14). Design principle: Hick's Law - quick decision making. Retrieved f from https://uxplanet.org/design-principles-hicks-law-quick-decision-making-3dcc1b1a0632.

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

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.

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.