A FRAMEWORK ANALYSIS OF DEEPFAKES: USING SWOT AND FMEA TO CALCULATE THE RISK POSED BY DEEPFAKES

By

Kastur Koul

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Computer Science

May 12, 2023

Nashville, Tennessee

Approved:

Douglas Fisher, Ph.D.

Pamela Wisniewski, Ph.D.

ACKNOWLEDGMENTS

I would like to thank Professor Doug Fisher, Kyle Moore, Jesse Roberts, and Samantha Bianco for their help and support. I would also like to thank my friends who were kind and patient enough to read my drafts.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Deepfakes have caused a mostly negative stir in the past five years. The first use of the term "deepfake" to refer to a video created by deep learning technology was in 2017, when a Reddit user by the name "deepfakes" posted a video which depicted the face of actress Gal Gadot on the body of an actress in a pornographic video (DHS, 2019), (Cole, 2017). The realistic nature of the video sparked concern in the media about the evolution of this kind of technology. Since then, there have been many other reported cases of the misuse of deepfake technology in politics and using the images of celebrities and popular figures, spreading fear and concern about deepfake technology and linking the term "deepfake" with negative connotations. Looking more critically at the media produced and considering the real-world consequences, published articles seem to mention the risks posed by deepfakes without using risk analysis techniques to give further details about why they are risks. In addition, the amount of negative press deepfakes receive overshadows the potential positive uses this technology has. This paper aims to analyze the potential risk posed by the current and future uses of deepfake technology and encourage the discussion of future uses of the technology.

This paper is sectioned as follows: first, I explain how deepfakes are created and detected using examples from current research. I then use the Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis method to list the aforementioned items for deepfakes. Next, I discuss the strengths and weaknesses of deepfakes in detail without assessing the risk posed as the strengths and weaknesses are intrinsic properties of deepfakes that do not warrant risk assessment. I conduct a Failure Mode and Effects Analysis (FMEA) style analysis of each of the opportunities and threats and assess the risk posed by each item as the uses brought about by opportunities and threats have effects on individuals and groups that can be analyzed. Finally, I average the scores assigned to the risk posed by each opportunity and threat and discuss what the scores mean. It is worth noting that this thesis focuses on deepfake videos and images, or more broadly visual deepfakes. There are other types of deepfakes that can be created such as audio and text, but this thesis will take a specific look at the creation, detection, and risks posed by video and image deepfakes.

CHAPTER 2

Preliminaries

2.1 Deepfakes Overview

The term "deepfake" is a portmanteau of the terms "deep learning" and "fake". A deepfake is a piece of digital media in which the face or movement of one actor is replicated on another actor using deep learning algorithms. What make deepfakes distinct from other photo and video editing techniques is the use of deep learning algorithms to create the illusion of the source image on the target image. There are three common kinds of deepfakes: head puppetry, face swapping, and lip synching, of which face swapping is the most commercialized and well known (Lyu, 2020). Head puppetry is when the source's head and upper shoulder movement is replicated on the target (Lyu, 2020). Face swapping is when the face of the source is swapped in for the face of the target while keeping the same facial expressions (Lyu, 2020). Lip synching is the manipulation of the lip region of the target to make it look like they are saying something they might not have said (Lyu, 2020). The creation, detection, and training of deepfakes can affect the environment. Deepfakes are an application of the broader deep learning field, which itself has effects on the environment. For example, one study from 2020 compared (among other factors) the carbon emissions by pounds generated by different deep learning methods applied to Natural Language Processing and found that the model can have up to 313,078 pounds of carbon emissions (Strubell et al., 2020). Researchers also found that the language generation model GPT-3 required 190,000 kilowatt-hours of electricity to train, which produced the same amount of carbon emissions as a car driving roughly 238,900 miles (Thompson et al., 2022). It would be reasonable to assume that deepfakes might have the same kind of environmental impact as other deep learning methods, but further research must be done to confirm this.

Google Trends notes that the interest in the search term "deepfake" began in late 2017, as seen in Figure 2.1. This corresponds to the time when Reddit user "deepfakes" posted their deepfake on Reddit. This deepfake was then exposed by Motherboard in December 2017 (Cole, 2017), sparking the concern surrounding deepfakes that continues to this day. The negative uses of deepfakes have gained attention in the media since 2017 and have spread concern and even fear among technology users. However, a lot of research has been done in detecting deepfakes, and some of this technology has been made available for people to use. In this chapter, I will summarize previous research on creation and detection to explain how the technology surrounding deepfakes works and provide a discussion of creation and detection in the future.

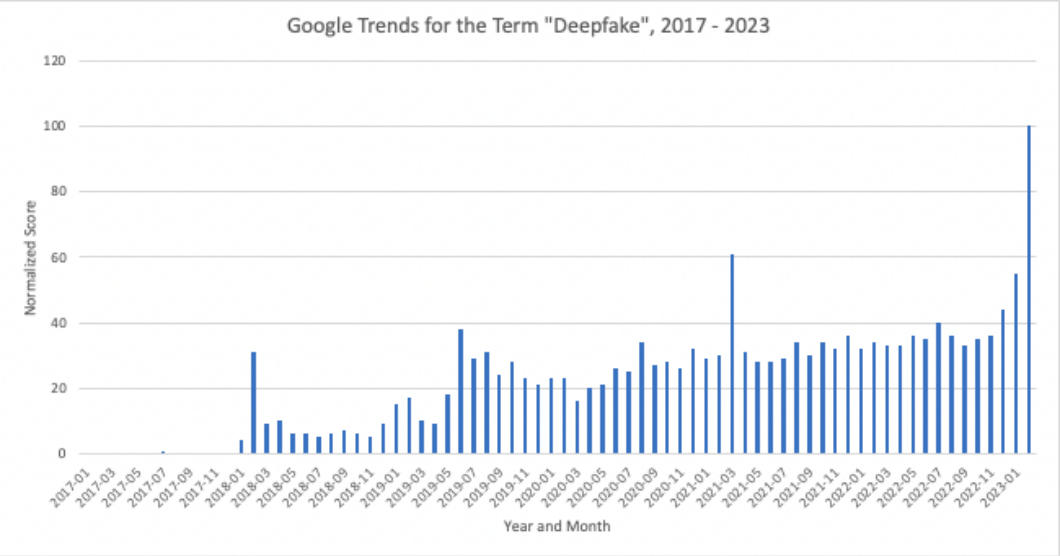

Figure 2.1: The trend of web searches for the term “deepfake” in the United States beginning January 1, 2017. Data adapted from Google Trends (Google, 2023).

2.2 Deepfake Creation

Deepfakes are created using deep learning techniques, making them a subset of the larger machine learning field. Commonly used techniques for creating deepfakes are Generative Adversarial Networks (GANs) and encoder/decoder networks. Yisroel Mirsky and Wenke Lee explain that GANs use a generator, which creates fake images, and a discriminator, which learns to tell the difference between the real and fake images (Mirsky and Lee, 2021). The generator creates images until its output cannot be distinguished from the original distribution and the discriminator is discarded (Mirsky and Lee, 2021). Apurva Gandhi and Shomik Jain give an example of the process of creating a deepfake in which common encoders and different decoders learn the features of a target, after which the common encoder and the target decoder are used to create a deepfake using a source image (Gandhi and Jain, 2020). An example of an encoder-decoder network can be seen in Figure 2.2. The following is a summary of some of the recent research done in deepfake creation and development in the related technology.

Hady Khalil and Shady Maged describe the use of autoencoders and decoders to create deepfakes. The autoencoder learns features from the source image and another encoder learns the features on the target (Khalil and Maged, 2021). The two encoders share their parameters, then generate a deepfake that is reconstructed with the decoder (Khalil and Maged, 2021). The size of the training dataset is directly related to how well this deepfake generation method worked (Khalil and Maged, 2021). The researchers then experiment with different algorithms to enhance deepfake images, noting that the use of the DFDNet algorithms generated the higher quality deepfake images (Khalil and Maged, 2021).

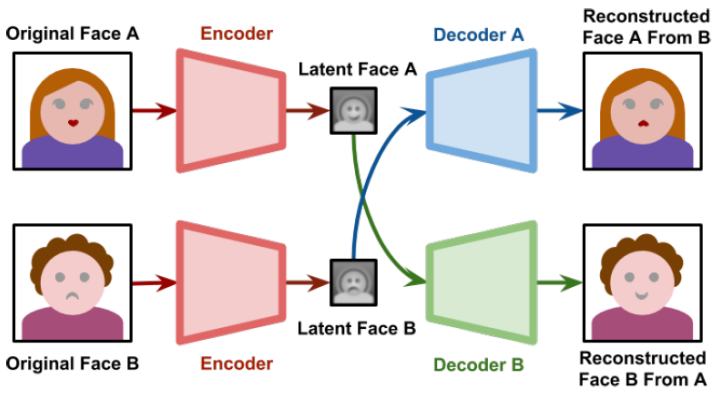

Fig. 2. Encoder –Decoder Network.

Figure 2.2: An example of an encoder-decoder network for creating deepfakes from (P and Sk, 2021)

Fan *et al.* use deepfake technology to create a real-time broadcast framework that will display a deepfaked face during a live stream. The researchers produce a closed-loop method of creating the real-time deepfakes using ResNet for face segmentation and feature extraction, HRNet for feature prediction, and a blended heatmap that is decoded to create the pictures (Fan et al., 2022). The result successfully hides the face of the human anchor (Fan et al., 2022).

Multiple papers discuss the same kinds of technology that create deepfakes without using the term "deepfake" at all. For example, in the landmark 2016 paper "Face2Face: Real-time Face Capture and Reenactment of RGB Videos", Thies *et al.* do not use the term deepfake in the paper as the term had not become well-known at the time of publishing (2016), but the paper discusses "facial reenactment" which can be classified as a kind of head puppetry (Thies et al., 2016).

Hariharan *et al.* discuss improving generated image quality using two GAN based algorithms. First, the researchers test the Deep Convolutional Generative Adversarial Networks (DCGANS) and the Style Generative Adversarial Network (StyleGANS) algorithms individually by having them generate images (Hariharan et al., 2022). Then they test a hybrid of the two algorithms and found that the hybrid produces better quality images than the two algorithms alone (Hariharan et al., 2022).

Prabhat *et al.* compare two GAN based algorithms to determine which can produce the highest quality images. In their paper, the researchers test the DCGAN algorithm and the Conditional Generative Adversarial Network (CGAN) by having the generator and discriminator train on the MNIST dataset (a dataset of handwritten digits popularly used in machine learning) (Prabhat et al., 2020). Performance analysis of the algorithms shows that DCGAN generated images were clearer but potentially takes longer to create than CGAN generated images (Prabhat et al., 2020).

2.2.1 Discussion

As with many pieces of technology, the creation of deepfakes will get easier with time. Currently, applications like FaceApp and FakeApp enable realistic face swapping in images and videos (Rana et al., 2022). There will be more applications available for smartphones and computers that will generate realistic deepfakes involving audio deepfakes and head puppetry. One feature that all future deepfake creation tools can include are digital fingerprints that can be picked up by detection tools. This way, the validity of a video or image would not be doubted and trust in both deepfake creation and detection tools can be created. Deepfake creation tools that offer their services for free can even include a logo as they do now. A current example of this can be seen in Figure 2.3. Deepfake creation tools that require payment for the use of their tools can also increase the price. The cost of the tool should not be a deterrent for a company wanting to use the tool but should put off individuals who do not have a comparable amount of money to spend. To avoid malicious behavior from individuals or groups with enough money to buy and/or use the tools themselves, the entities creating deepfakes can instead offer the service for hire. This way, deepfake creators can mitigate potential risks by monitoring the inputs and outputs of the deepfake creation tools.

Figure 2.3: (left) A still from a video created by the application FaceMagic, which uses face-swapping to creating videos. In this still, my face has been swapped with actress Audrey Hepburn's in the movie *Roman Holiday* (1953). The FaceMagic logo can be seen in the top left. (right) A still from the same scene in *Roman Holiday* with Hepburn's actual face for comparison (IMDb, 2023).

The creation of deepfakes should not be a cause of concern in the future. The technology should be managed in a way that does not cause distrust yet creates realistic and entertaining ways for people to engage in the technology. The consequences and calculated risk of the effects of creating a deepfake depend on the

use of the final product. The only way that deepfakes would not be created and circulated is if they are banned by the federal government. As deepfakes are created and circulated currently, it is almost guaranteed that they will continue to do so in the future. Detection is the counterpart to the creation of deepfakes, and it certain to happen for some uses but not necessary for others depending on how severe the consequences of the use are.

The technology that goes into deepfake creation is not new. Recent papers focus on the various uses of the technology and potential improvements to the image quality produced by the algorithms. As the research continues to develop better deepfakes, the concern surrounding them continues to grow. To potentially mitigate this concern, researchers have also focused on deepfake detection methods.

2.3 Deepfake Detection

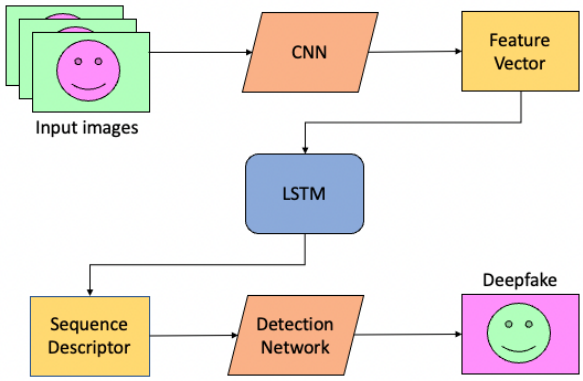

Figure 2.4: Deepfake detection using CNN and LSTM based on (Chadha et al., 2021)

Recent research indicates that the most used methods of deepfake detection are a combination of convolutional neural networks (CNNs) and long short-term memory networks (LSTMs). Algorithms look for inconsistencies in physical/physiological aspects or signal-level artifacts, or they can be data-driven methods that use various deep neural networks trained on real videos and deepfakes (Lyu, 2020). A literature review done in 2022 indicated that CNN models have the highest percentage of use among deepfake detection models (Rana et al., 2022). An example of a CNN-LSTM network can be seen in Figure 2.4. Most research as of 2022 uses detection accuracy to test the effectiveness the proposed detection methods (Rana et al., 2022). The following is a summary of some of the research done in deepfake detection in recent years.

Guera and Delp propose the use of a CNN and LSTM for deepfake detection. The CNN is used to extract features at a frame-by-frame level. The features are then fed into a recurrent neural network that learns how to classify real or deepfaked videos, and the LSTM is used for temporal sequence analysis. The system was able to detect a deepfaked video "with as few as 2 seconds of video data" with 97%+ accuracy (Güera and Delp, 2018). In a later research paper, Yunes Al-Dhabi *et al.* propose another combination of CNN and RNN

for detection, followed by the use of LSTM to check for accuracy. This algorithm correctly distinguishes between real videos and deepfaked videos with an accuracy of 99.8% (Al-Dhabi and Zhang, 2021).

Jiwtode *et al.* propose a method to identify deepfakes that uses InceptionResnetV2 CNN to perform a frame-by-frame detection of anomalies in the video that could be used to identify it as a deepfake. The neural network detector has a high precision with real time data (high true positive rate, with a comparatively low false-negative and false-positive rates) (Jiwtode et al., 2022).

Koopman *et al.* discuss the use of photo response non uniformity (PRNU) analysis as a deepfake detection method. The input video is turned into several frames which are then cropped and grouped together. PRNU patterns are created for each of the frame groups and are compared to each other, returning a normalized cross correlation score (Koopman et al., 2018). Statistical analysis of the scores showed a correlation between the mean correlation scores and the video's authenticity (Koopman et al., 2018). The method reports a 3.8% false positive rate and a 0% false negative rate on the small dataset (10 real videos between 20 to 40 seconds long and 16 deepfakes) used for testing, but the author suggests testing on larger datasets to confirm (Koopman et al., 2018).

Ciftci *et al.* propose the use of biological signals (changes of color and motion for example, specifically using photoplethysmography [PPG]) as indicators of deepfakes. The researchers create a video classifier based on physiological changes and encapsulated those the signals in PPG maps to allow development of a CNN-based classifier. The detector has a high rate of accurate, having 91.07% accuracy in wild video classification and 96% accuracy in constrained video classification accuracy (Ciftci et al., 2020).

Outside of algorithmic techniques of detecting deepfakes, the human eye can spot inconsistencies in human-like images and videos that would indicate that a video is fake. (Groh et al., 2021) discusses two online studies done in which human participants and a deepfake detection model were asked to identify real and deepfaked videos. The study found that participants were able to identify deepfakes without the aid of detection, even outperforming the detection software at times. It is mentioned in (Groh et al., 2021) that there is a region of the human brain which specializes in detecting faces, meaning that people can easily detect visible digital artifacts like inconsistency in the size of the features to the face or mismatched timing in the source layout's movement on top of the target. If deepfakes cannot get the features to look or behave right on the face, the entire video is easily debunked. For example, putting features that clearly belong on a larger face on a much smaller one makes the video look incorrect. An example of this can be seen in Figure 2.3, in which the still on the left is pulled from a longer video.

An interesting study in deepfake detection avoidance comes from Gandhi and Jain. They create "adversarial perturbations" to enhance deepfakes created by GANs to make them harder to detect using common deepfake detectors. The researchers explain that Deep Neural Networks and other such models are vulnera-

ble to “input data that has been perturbed to make the model misclassify the input” (Gandhi and Jain, 2020). The perturbations to the deepfaked images are small enough that the perturbed images are harder to distinguish from the unaltered images. Gandhi and Jain successfully make the deepfakes harder to distinguish with common detection methods.

2.3.1 Discussion

As deepfake technology becomes increasingly used, detection methods would also find an uptick in usage. There would most likely be widespread use of CNNs built into video and image tools that could help users identify videos. For example, a smartphone and computer application could be built that uses CNN and LSTM to analyze videos and images to determine if they are authentic. Videos and images that are clearly labeled as “deepfakes” or come from reputable sources do not necessarily need to be checked with this application. Users can also indicate if they know that the video is real or not, providing training feedback to the application as it learns. There is a potential for bad faith users to trick the algorithm by indicating a deepfake video is real. In this kind of instance, backup checkers can be hired (humans) or implemented (algorithms). The main problem with such an application is that it would have to monitor the user’s activity on various applications, but because there is no personal data being collected there is no threat to the user’s privacy. The technology for this kind of application already exists. (Vamsi et al., 2022) states that the proposed method, a combination of a CNN and LSTM, and could be integrated into applications that could circulate deepfakes. Companies that produce the deepfake sharing application could also implement their own tools for detecting deepfakes. It would not be very difficult for these companies that employ thousands of software engineers to create tools that check through videos and images and flag warnings for deepfakes. An example of this kind of combing and tagging comes from Meta in recent years. During the COVID-19 pandemic Facebook implemented a product wide tool that would tag any post that touched the topic of the pandemic, vaccines, and/or the virus with a note of caution about the content. An example can be seen in Figure 2.5. There is also the chance that deepfakes that are shown to be fake might cause people to think that they are better at spotting deepfakes than they actually are if they guessed the fake before official detection was done, which can lead to overreliance on human eyes alone for detection. Human eyes alone will eventually not work in spotting deepfakes with the improvements in creation technology and would create a distrust in any forms of visual digital media.

The proliferation of detection tools might lower the potential risk posed by other future uses of deepfakes because as the technology spreads and improves, deepfakes might become easier to find using detection algorithms. In the two-year period between 2018 and 2020, there were over 100 studies done on deepfake detection (Rana et al., 2022). This should be a reassuring fact for those concerned with the proliferation of deepfakes online as it demonstrates that the concerns Internet users have are being addressed through research

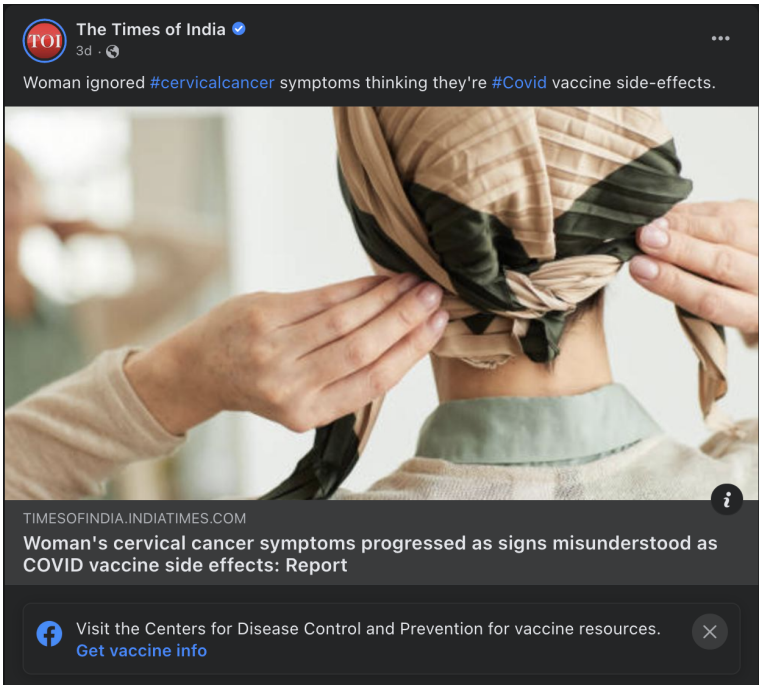

Figure 2.5: A screenshot of a Facebook post with a COVID-19 related flag

in detection methods.

2.4 Previous Risk Analysis Work

There are few works that just analyze and discuss the risk posed by deepfakes. One of these few comes from Ali *et al.*, who discuss the potential malicious uses of deepfakes especially in politics (Ali et al., 2022). The article does not use a framework to analyze the risk posed by these malicious uses but does list various digital artifacts that can be used to detect deepfakes with the human eye. (Ali et al., 2022) also briefly states that deepfakes are "primarily just a technical tool with more positive uses than negative ones" but does not elaborate on the positive uses.

Gamage *et al.* use two research questions to guide their study of deepfakes, analyzing the conversations Reddit communities have about deepfakes and the societal implications of these conversations (Gamage et al., 2022). The researchers discuss at the end if deepfakes are concerning, and according to their study there is reason for concern. They do point out that deepfake technology is "a double-edged sword" that has many positive uses that should not be eliminated (Gamage et al., 2022).

A study done by Pew Research Center in 2019 shows that adults in the United States hold generally negative views about altered images and videos. The study states that 77% of adults in the United States say that "steps should be taken to restrict altered videos and images that are intended to mislead", with 22% being

in favor of accessing and pushing these videos and images (Gottfried, 2019). This concern is attributed to the fact that altered videos and images have a far reach, with 66% of the survey participants sometimes seeing and 15% often seeing altered videos and images that are intended to mislead (Gottfried, 2019).

This thesis provides a novel approach to the analysis of deepfakes by using two analysis frameworks to calculate the potential risk posed by deepfakes. It is also novel because it breaks deepfakes down into the strengths, weaknesses, opportunities, and threats posed and discusses each one, which previous analyses do not do.

CHAPTER 3

SWOT Analysis of Deepfakes

In this chapter, I will conduct a SWOT analysis of deepfakes. SWOT is an analysis method used generally for strategic planning in which the strengths (S), weaknesses (W), opportunities (O), and threats (T) are identified in the internal and external environments of an organization (Kharchenko et al., 2022). Strengths can be defined as something that the system does well, while weaknesses hinder the system's performance (Ahmed and Kumar, 2022). Opportunities and threats are variables that come from the environment outside a system that can either positively or negatively impact the system (Ahmed and Kumar, 2022). While more commonly associated with applications in business, this analysis technique can be used to evaluate the various strengths, weaknesses, opportunities, and threats for systems and technologies outside of business. This includes examples such as the effectiveness of online education (A. Safonov et al., 2021), methods for evaluating a system's fit to its real-world environment (Hertzum et al., 2023), and various distributed generation methods for power systems (Ahmed and Kumar, 2022). The key benefit of using the SWOT system is that it accounts for both the positives and the negatives of a given system or technology. A weakness of this analysis method is that it does not provide in-depth analysis or diagnosis of the listed items (Minsky and Aron, 2021). However, this weakness is mitigated by further discussion of the listed strengths, weaknesses, opportunities, and threats.

For this analysis, I will list the strengths, weaknesses, opportunities, and threats of deepfakes in Table 3.1. As SWOT is used generally for business systems, different definitions are needed to discuss the SWOT items for deepfakes. These new definitions will be defined in terms of how the deepfakes are used and experienced by the end users. Strengths for deepfakes are the currently exhibited positive attributes the technology displays in its use by an end user. Weaknesses for deepfakes are the currently exhibited negative attributes of the technology that might hinder the experience for end users. Opportunities for deepfakes are the potential positive uses of the technology in various fields. Threats of deepfakes are the potential negative uses in various fields. In total, there are 3 strengths, 3 weaknesses, 7 opportunities, and 7 threats listed for deepfakes.

Strengths (S)	Weaknesses (W)	Opportunities (O)	Threats (T)
Saves time and money for small companies as they can be cheaply made. Uses deep learning techniques Innovation in digital realism.	Most deepfakes have detectable differences that make them too easy to spot. Uncanny Valley effect can spoil the experience of using a deepfake. Availability and quality of training data might not be enough to make a deepfake of any random person.	Use in the Fashion industry to introduce accessibility Use in Entertainment to recreate popular actors' faces and accurate lip dubbing. Use in Education to create interactive learning tools. Use in AR/VR to create immersive experiences. Use in Video Game Development to assist in the development process. Create opportunities for Trustworthy AI by being transparent, acessible, and diverse. Embodied chatbots for telehealth and teletherapy.	Use in social engineering through impersonation Ethical concerns about the creation of deepfakes using someone's image without their consent. Difficult to regulate and use of deepfakes as evidence in cases of law. Online harassment (such as blackmail and impersonation) resulting in a lack of privacy and security. Use of deepfakes to create nonconsensual porn. Spread of misinformation to bolster personal agendas. Use of deepfakes in politics.
Total: 3	Total: 3	Total: 7	Total: 7

Table 3.1: SWOT analysis of deepfakes

CHAPTER 4

Strengths and Weaknesses

This chapter will discuss the strengths and weaknesses of deepfakes as itemized by the SWOT analysis in Chapter 3. Uses of deepfakes will be discussed in order to provide examples of how the strengths and weaknesses present themselves. These uses will be further analyzed in Chapter 5. As the strengths and weaknesses are defined as attributes, which are inherent properties of deepfakes that appear in the various uses they are put to, a risk analysis will not be conducted in this chapter.

4.1 Strengths

The strengths of deepfakes are the positive attributes of the technology that appear in its various applications. The following is a discussion of the strengths of deepfakes listed in Table 3.1.

Deepfakes are created using deep learning techniques, as discussed in Chapter 2. Many if not all of the research papers discussing the creation and detection of deepfakes use deep learning algorithms such as GANs and CNNs in their work. The outputs of these algorithms are pictured in the papers, many of which report success in using the deepfakes through the high quality of the outputs. The deep learning algorithms also create tools that can be easy to use, an example of which can be found in (Fan et al., 2022).

Any innovation in technology can potentially save time and money for the people using it. Deepfakes are one such piece of technology. The deep learning models used to create deepfakes can potentially run faster and cost less than a human doing the same job. Small companies and start-ups can benefit from the time and money saved through the use of deepfakes. Instead of needing to hire real models or actors for marketing or training videos, which would require paying the person and several takes that take time, small companies can use tools like Synthesia that create deepfake models cheaper and have them do the job faster. Using deepfakes means that the money and time spent on one task can now go towards another that might need the resources more or can even be donated. The time and money saved can provide beneficial uses that help the individual or the company in the long run. There is a high chance that deepfakes will be used to save time and money in the future, especially for small businesses.

Realism is the feature of deepfakes that speaks most to their quality. Some of the best deepfakes are the ones that look so realistic that they easily fool a human viewer. Realism is needed in various applications of deepfakes. In entertainment, for example, the use of realistic deepfakes can recapture the image of deceased celebrities and can de-age older actors. Realistic deepfakes can also be used to populate crowds in the background of scenes in movies and television shows, saving production time and money without breaking the

illusion for audiences. However, malicious uses of realistic deepfakes can cause severe emotional, financial, or societal damage to the person or people depicted in the video or the intended audience of the deepfake. The use of realistic deepfakes in cases such as deepfake pornography (discussed further in Chapter 5.3.5) can potentially damage the reputation of the person depicted in the video. In politics (discussed in Chapter 5.3.7), realistic deepfakes, if not detected in time, can lead to outbreaks of war in the worst case which does many people a great amount of emotional and financial harm.

4.2 Weakness

The weaknesses of deepfakes are the negative attributes of the technology that appear in its various applications. The following is a discussion of the strengths of deepfakes listed in Table 3.1.

Some deepfakes are very believable at first glance and can take multiple views or algorithmic detection tools to uncover as fake. Others can be instantly detected by the human eye alone. For video deepfakes, it is possible to see odd mouth movement that does not match the audio or mismatched facial features on a face that is too large or too small. Deepfake images generators might suffer from poor training data that can result in output that is blurry or disproportioned. This is where the Uncanny Valley comes in. The Uncanny Valley, described first by Professor Masahiro Mori in 1970 and visualized in Figure 4.1, describes the increasing affinity people have for human-looking robots until that affinity comes to a valley (Mori et al., 2012). Deepfakes, with their current generally off-kilter outputs, fall into the uncanny valley. Irregular eye and mouth movement and audio that does not match the avatar's movement are two examples of deepfake features that cause the deepfakes to fall into the uncanny valley. Some deepfakes make it out of the Uncanny Valley with the use of advanced GANs and special effects technology. An example of this is DeepTomCruise, a TikTok account that posts very realistic face-swapping videos that are created with machine learning algorithms and special effects software (Vincent, 2021).

As with any machine learning algorithm, deepfake creation and detection algorithms require a lot of training data in order to produce an output. The quality of the training data can affect the quality of the output as well. For deepfakes, should the creation algorithms be trained on blurry, pixelated, or otherwise low-quality images, it will produce a low-quality output. Similarly, if the algorithm is not fed enough input data to learn from, the output deepfake will not be very convincing as it might be missing crucial facial angles to learn from. For example, one cannot guarantee what the front of someone's face looks like just by looking at that person's profile. The key characteristic of the training data for deepfakes is that the data is of human faces. There are two key issues with needing human faces to train the deepfake models. The first is the number of images needed to create a convincing deepfake of the person. Common databases used in deepfake research contain anywhere from as few as 49 deepfake videos to as many as 420,053 images

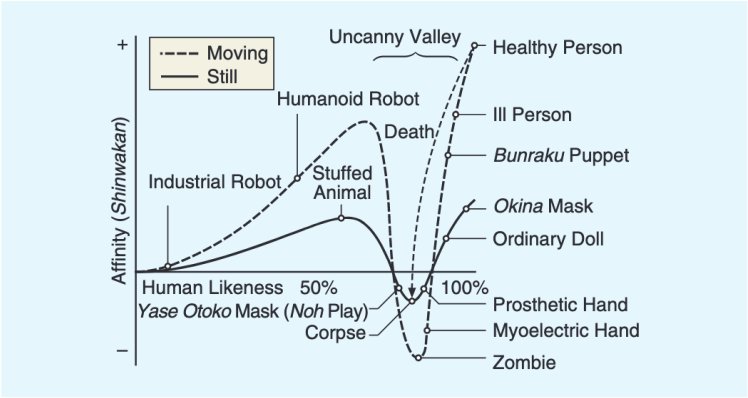

Figure 2. The presence of movement steepens the slopes of the uncanny valley. The arrow's path represents the sudden death of a healthy person. [Translators' note: *Noh* is a traditional Japanese form of musical theater dating to the 14th century in which actors commonly wear masks. The *yase otoko* mask bears the face of an emaciated man and represents a ghost from hell. The *okina* mask represents an old man.]

Figure 4.1: The Uncanny Valley from (Mori et al., 2012)

(Rana et al., 2022). It can be assumed that the best performing deepfake algorithms require as large and as diverse a dataset as possible to train on, and therefore creating a quality deepfake of a specific person would require several thousands of pictures of their face and shoulders from multiple angles in good lighting. Most people do not have that many pictures of themselves readily and easily available, so it would be difficult for extremely realistic deepfakes to be made of them. Celebrities and models, on the other hand, do have these kinds of pictures readily available and are used often as training samples. There is already a dataset available that is made entirely of celebrity faces (Li et al., 2020). The use of celebrity images in deepfakes is discussed further in the next chapter. The second issue is the consent of the person whose image is being used. This will also be discussed in the next chapter, but to briefly address it here, the nonconsensual use of a person's image for commercial purposes can be considered illegal. However, training data should not be sold or used for commercial purposes. While people might be uncomfortable with their image being used to train the deepfake model, their image as part of the dataset is safe so long as the dataset is not sold. Ideally, people should give their written consent for their image to be used in training data for deepfakes. This would provide peace of mind and liability for the researchers and developers using the image and the people providing the image to be used.

CHAPTER 5

Opportunities and Threats

This chapter will discuss the opportunities and threats of deepfakes as listed in the SWOT analysis from Chapter 3. The opportunities and threats of deepfakes can be seen in their many uses, and so a risk analysis will be conducted for each of the opportunities and threats discussed in this chapter.

5.1 FMEA Risk Analysis

The risk analysis of the opportunities and threats will be done using an FMEA style analysis. FMEA is a method used to identify and address potential problems and the effects on the system (CMS, 2021) and provides an insight into what the behavior of a system will be given a single point of failure (Shi et al., 2011). This analysis process produces a Risk Priority Number (RPN), which can be calculated by finding the product of the calculated severity, occurrence, and detection of a given problem in a system (Stanojević and Ćirović, 2020). FMEA can be adapted in different ways. One example of this is from the FMEA guide from the Centers for Medicare & Medicaid Services, which does not include the calculation of the RPN (CMS, 2021). A notable weakness of the FMEA method is that the three values that make up the RPN do not necessarily have equal significance in terms of risk (Stanojević and Ćirović, 2020). For example, a problem of low risk but high occurrence might make the RPN higher than the actual risk posed. Despite this, FMEA is a widely accepted method of risk analysis (Stanojević and Ćirović, 2020). The mentioned weakness could be viewed as a strength because it encourages solutions for every problem regardless of severity, which could result in a better product or system. Other strengths include its cost effectiveness (Goddard, 1993) and consideration of every part of a system. This later point is exemplified in the FMEA Reference Manual by the Chrysler Motor Company, where it states "the FMEA discipline requires a Design FMEA for all new parts, changed parts, and carryover parts in new applications or environments" (Chrysler Corporation, 1995). The FMEA method has been tested by time, with papers using FMEA to analyze software appearing as early as 1983 (Goddard, 2000) and the method itself being used since the mid 1960s (Chrysler Corporation, 1995). Some examples of its use include validation of embedded real-time systems (Goddard, 1993), analysis software architecture design (Kim, 2014), and analysis of automotive design (Chrysler Corporation, 1995).

For the risk analysis of deepfakes, I will adapt FMEA to use the severity, occurrence (referred to in the rest of the thesis as the probability), and detection to calculate the expected risk of the deepfake use. The FMEA analysis will be applied to the opportunities and threats of deepfakes listed in Table 3.1 as these are the uses that pose risks. Strengths and weaknesses are inherent parts of deepfakes that show themselves regardless of

use, and it is the exploitation of the weaknesses (threats) or the capitalization of the strengths (opportunities) that creates the risk to be analyzed.

I will estimate the probability of the use (how likely is the use of the deepfake to occur) on a scale of 0.2 to 1, with 0.2 being most unlikely to happen and 1 being guaranteed to happen. Table 5.1 gives the definition of each probability score. As a note, the probability scale does not start at 0 because it is highly likely that deepfakes will continue to be produced and circulated as they are currently, and this thesis will not be considering non-creation of various uses. It is also worth acknowledging that this metric does not consider a difference in scope in terms of how many people are affected. Deepfakes can have effects on and be used by an individual or a large group depending, and future studies can improve upon the scale proposed here by considering the scope.

Probability	**Description**
0.2	Not likely; Most likely the use will not occur in the future.
0.4	A little likely; The use might happen, but the chance is unlikely.
0.6	Likely; The use might happen.
0.8	Very likely; High chance the use will happen.
1	Guaranteed; This use will happen in the future.

Table 5.1: Probability Scale

There has been research done in deepfake detection with the use of deep learning techniques, and some deepfakes that get spread today can be detected by the human eye because of visual digital artifacts. I will score deepfake detection for each of the uses on a scale from 0 to 1 with 0 meaning that the deepfake is impossible to detect and 1 meaning that the deepfake is guaranteed to be detected, as displayed by Table 5.2. This score will then be used with the probability and severity score to determine the expected risk.

Detection	**Description**
0	Impossible to detect; The deepfake will not be caught.
0.2	Difficult to detect; The deepfake very likely not get caught.
0.4	A little difficult to detect; The deepfake might not get caught.
0.6	A little easy to detect; the deepfake might be caught.
0.8	Easy to detect; The deepfake will very likely be caught.
1	Guaranteed detection; The deepfake will be caught.

Table 5.2: Detection Scale

As a note, the detection score does not account for two factors: time to detection and prevalence of detection. The timing in which deepfakes get detected is important. For deepfakes that pose severe consequences, lack of timely detection can have extremely severe consequences for individuals and society. However, adding time as a factor of detection complicates the scoring as time can fall into a probability distribution where some deepfakes can be detected right away and others will be detected over time. The prevalence of detection is

another factor that will not be included in the detection measure. While detection tools are available and there is research being done into improving deepfake detection, it is not easy to determine how commonly these tools are being used. Further, there are possible distinctions between detection by individuals compared to whole groups and detection by experts versus nonexperts. I acknowledge that time and prevalence are important factors to consider, but they would complicate the evaluation of detection in this thesis. They would be factors to include in further studies of deepfakes and their impacts.

I will analyze the potential severity of each item using the consequentialist ethical framework. Consequentialism is an ethical framework similar to utilitarianism that determines the moral right of an action based on the consequences of that action (Sinnott-Armstrong, 2022). For this analysis, the consequences will be defined as the positive or negative emotional, financial, or societal effects on an individual caused by the specified opportunity or threat of a deepfake. There will be a severity score assigned to both undetected deepfake use and detected deepfake use. I will score the potential severity of the effects of the proposed use on a scale from -4 to 4. Increasing negative values indicate increasingly negative consequences of the opportunity or threat. The same applies in reverse for positive values. Positive values indicate benefits of the opportunity or threat, and increasing positive value indicates more beneficial applications. Table 5.3 illustrates the definition of each severity score.

Severity	Description
-4	Extremely severe; Irreparable and guaranteed negative emotional, financial, or societal damage.
-3	Very severe; Guaranteed negative emotional, financial, or societal damage that is potentially but not easily reparable with time.
-2	Severe; Negative emotional, financial, or societal damage that is reparable with time.
-1	A little severe; Mild negative emotional, financial, or societal damage that is easily reparable.
0	Neutral; Neither positive nor negative consequences.
1	Few benefits; Mild positive emotional, financial, or societal outcomes.
2	Some benefits; Positive emotional, financial, or societal outcomes
3	Many potential benefits; Long lasting positive emotional, financial, or societal outcomes.
4	Guaranteed benefits; Life-changing positive emotional, financial, or societal outcomes

Table 5.3: Severity Scale

Finally, the three scores used to calculate the expected severity of the risk of the deepfake use. This will be done by calculating the weighted average using the following calculation:

$$RPN = P * ((D * S_D) + ((1 - D) * S_U))$$

Where P is the probability of the use occurring, D is the likelihood that the use is detected, S_D is the severity of the use if the use is detected or known, S_U is the severity of the use if undetected, and the RPN is the expected risk of the use. The worst RPN is a -4, which means the use is guaranteed to occur, is impossible

to detect, and has extremely severe consequences. The best RPN is 4, which is a use that is guaranteed to occur, will easily be detected, and has guaranteed benefits. A future complete risk analysis can include the severity posed by the non-creation of deepfakes for a given use, but for this analysis I will discuss the risk posed by the use of deepfakes.

It is also worth mentioning that advances in technology and subsequent automation of jobs potentially result in the loss of jobs. For example, the job of a stock trader on Wall Street became obsolete with the introduction of algorithms that did the job faster, cheaper, and more precisely. Deepfake use in different fields can be considered a form of automation as it would replace the need for human workers. However, that automation does not necessarily mean a decrease in the number of jobs available. For example, in 2021 the World Economic Forum predicted that automation would result in a net increase of 58 million jobs (Hanspal, 2021). Because automation has the potential to create or remove jobs, I will not factor potential job loss into the risk analysis but will discuss what the effects might be where applicable. It is important to acknowledge that the effect on the job market is a fact to be considered for any evolutionary step in the use of deepfakes.

5.2 Opportunities

There are a variety of opportunities created by the existence of deepfakes, many of which build on the strengths. In this section, I will discuss some opportunities for the use of deepfakes in various fields and conduct an FMEA analysis of the risk posed by these opportunities.

5.2.1 Fashion

As GANs are becoming better at producing full body images, there is a use for the technology in the future of the fashion industry. Shoppers can use GAN-made images of themselves to virtually try on clothes instead of having to go to the store. This would be a low severity application of deepfake technology as it aims to provide a useful service. Deepfakes in the fashion industry would open fashion up to accessibility. Customers who are unable to or find it difficult to try on clothes in-person in the stores would have the option to try it on online. Having this technology would not eliminate the option of going to the store physically but would provide another way of achieving the same end. People would start saving money as they would not be unnecessarily spending, the number of returns would reduce, and travel time would be saved as people would not need to go to the store. Reducing travel time would also be beneficial for the environment as less people traveling means less cars on the roads. A concern for this kind of use of deepfakes would be hacking. Should the application be hacked, it poses the worst-case scenario of becoming a deepfake pornography application with access to the likeness of many people. The right kind of security could provide mitigation methods for hacking, but it is still a viable concern.

Deepfakes in fashion offer a huge step in accessibility for the fashion industry as it taps into a market of people who might not have the ability or opportunity to shop with ease. However, the threat of hacking and losing sensitive personal information adds a level of severity to this application of deepfakes. In this case, the consequences of losing the personal information add a level of risk to the use because loss of personal information can cause emotional and financial damage for the target whose information is used. Detected and known deepfakes in fashion can be put at a severity level of 2, erring on the side of caution because of the personal information involved. The severity of the potential loss of personal data should the deepfake go undetected would be -2 on the scale. For positive uses, there would be little need for detection as it is known that deepfakes are being used, but for negative uses detection becomes important for the protection of the data and by extension the identity of any person whose images are used to create the deepfakes. There are some ways in which detection can be applied. One idea would be to have a detector that operates similarly to web crawlers to see if the deepfake models had been used anywhere other than where they were created for. Use of detectors like this would make the detection score 0.6, where the crawling detectors and other guards would make it slightly easier to find misuse of these deepfakes. It would be likely that most major clothing and department stores would use this technology in the future because the increase in profits from the new market would be too great to ignore, putting this at a level 0.8 on the probability scale. Overall, deepfakes in fashion earn an RPN of 0.32.

5.2.2 Entertainment

There are multiple uses for deepfakes in the entertainment industry. Deepfakes could be used to create more realistic dubbing for movies and television shows. The use of audio deepfakes, head puppetry, and lip synching can be used to create accurate dubbing for movies and television, making the artform more accessible to audiences in different countries. An example of this comes from (Suwajanakorn et al., 2017), in which researchers produced a high-quality video of President Obama using recurrent neural networks and LSTM. Deepfake technology could also be used to preserve the likeness of popular actors so that they can star in movies in the future long after they have retired and earn some money from the licensing of their image. This would be a great bit of technology for the entertainment industry to have because once they can get the likeness licensed for use, they have the potential to draw in high revenue for the movies in the future. This technology has already been used in the entertainment industry. Famous examples include the likeness of a younger Mark Hamill used in the show *The Mandalorian*. To fit the timeline of the series, a deepfake version of a younger Mark Hamill playing Luke Skywalker was created "with images of young Luke overlaid over Hamill and [stand-in actor] [Max] Lloyd-Jones on set" where the images were pulled from old pictures and footage (Hunt, 2022). It is very likely that this technology will continue to be used in the future because of

how easily available pictures and videos of popular actors are. It is worth touching on the effects of deepfakes on the job market in the entertainment industry. Instead of having their faces be on the screen, actors might find themselves being the stand-in actor whose face will be swapped with a popular actor's as done in *The Mandalorian* and DeepTomCruise. Deepfakes will likely not replace the physical actor because a virtual person cannot interact with the physical environment. Physical contact between actors can be a big part of the relationship between characters in the movie or show, so the role of an actor might change a little. The only way deepfakes will completely replace actors is if the entire production is done on a computer, meaning the backgrounds, props, and actors are all computer generated. The need for voice actors translating for characters might no longer be needed with the accurate dubbing that can come from lip synching deepfakes.

To briefly discuss the application of United States law, an actor's name, image, likeness, and voice are protected because they manage and monetize their image as part of their livelihood. Actors would have the choice to allow the use of their image created by a deepfake. Should they refuse to license, and their image or voice is still used in a deepfake for commercial use, the actors would have the ability to sue because there are regulations in place that would protect a celebrity's image. In the United States, sixteen states and the District of Columbia have definitions of the Right of Publicity (Project, 2022c). In California, the Right of Publicity protects a person's photograph, likeness, and voice, and violation of the law "entitles the plaintiff to the 'actual damages suffered'…" (Project, 2022a), where damages are the negative consequences inflicted on the target due to the spread of the deepfake. The consequences of using a person's likeness without permission is enough to deter misuse of the technology in the entertainment industry. In New York, the Right of Publicity similarly protects a person's name, portrait, picture, and voice (Project, 2022b). Generally, the use of a person's name, likeness, voice, or any other personal attribute for an exploitative purpose without their consent is grounds for a lawsuit in the United States (Project, 2022d).

The known use of deepfakes in entertainment is not a worrisome problem because it would require the permission of the owner of the likeness being used, and this likeness is protected by law in many states. Potentially negative risk can be mitigated by detection methods and the law. Since it would be known that deepfakes are being used (hence they are "detected"), deepfakes in entertainment would rank at 3 on the scale of severity because the use of deepfakes has the potential to change the way the entertainment industry creates movies. There might be some cases in which a clever programmer could get away with the use of a deepfaked image of a celebrity for a negative purpose. However, because celebrities are well known figures, it would not take very much time for undetected uses to be detected. Because of this, the severity of an undetected deepfake would be -1. Because the technology already exists and big-name celebrities generate revenue for movies, it is very likely that deepfakes will be used in entertainment in the future. Given this, deepfakes in entertainment can be ranked as 0.8 on the probability scale. Finally, visual detection of deepfakes might

actually be detrimental to their use in entertainment as visual artifacts would betray a lack of quality and effort put into producing the show and would pull audiences out of the experience by distracting them from the story. Deepfakes that use a celebrity's image in selling a product or narrative would be reported quickly through social media and/or news outlets. Therefore, deepfakes in entertainment earn a detection score of 0.8. Overall, deepfakes in entertainment earn an RPN of 1.76.

5.2.3 Education

Deepfakes could have a practical use in education. Deepfake tools could enable students to talk to famous historical figures, creating memorable learning opportunities for students. Teachers could set up virtual meetings with historical figures relevant to the subject being taught, such as a meeting with Abraham Lincoln where he teaches about the American Civil War. Students could read William Shakespeare's plays with the Bard himself or learn about the Laws of Motion with Isaac Newton. The deepfake technology here would combine audio, head puppetry, and learning facial expression, and would tap into other artificial intelligence tools like GPT. Professor Ole Molvig, Assistant Professor of History at Vanderbilt, has an example of this kind of technology. Talk To Einstein, created by Prof. Molvig in collaboration with resemble.ai, is "an experiment in algorithmically generated (or 'synthetic') media (AGM)" in which AI algorithms are used to recreate the voice, image, and written style of Albert Einstein (Molvig, 2023). Projects like Talk to Einstein would be great examples of the use of deepfakes in education. They would be low-risk tools for students and teachers as the information and technology would be provided by a separate service and would be widely used because it has the potential to improve the quality of education in K-12 through immersive and memorable experiences.

Most tools for educational purposes are low risk because they operate under limits. Certain features are unavailable for academic users of commercial applications. It would be reasonable to assume that deepfake technology used in academic settings would operate under similar limits. There would be no need for personal data collection because the deepfake would be an educational tool, and revenue can be generated by licensing the use of the product. As it is known that the tool uses deepfakes and can be considered detected, it is fair to assume that deepfakes used in educational setting rank 3 on the severity scale because there are little to no negative consequences of the use and could be positively beneficial for students. Undetected use of the deepfakes in education could be damaging depending on how they are used. The spread of misinformation or heavily biased information through the use of deepfakes can have damaging consequences for students learning from them until the information is otherwise proven false (which might still not work should the student choose to reject the new information). On the other hand, undetected use of deepfakes that spread accurate information might be fine if left undetected. To err on the side of caution, the severity of undetected deepfakes can be marked at -2. The use of deepfakes in education would be widespread in school districts

throughout the United States, but their spread depends on affordability and availability of the technology. Therefore, deepfakes in education would be a 0.6 on the probability scale. In terms of detection itself, the people being recreated with the deepfake technology would already be dead so it would be clear that the deepfake person is not real. However, the earlier mentioned example of misinformation spread in education does warrant detection put in place by the creators to ensure a quality product for education. The combination of the positives and negatives make the detection score 0.6. This gives deepfakes in education an RPN of 0.6.

5.2.4 Augmented/Virtual Reality (AR/VR)

Deepfakes can be used in augmented reality (AR) and virtual reality (VR). VR is the use of digital technology to create completely virtual, interactive three-dimensional experiences. AR is a subset of VR that combines the real and virtual. The process of creating virtual models of real objects in AR and VR is similar to the process of creating deepfakes. Multiple high-quality pictures from various angles and in good lighting need to be fed into a model that can recreate the object in a virtual setting. Deepfakes can be used to create immersive AR/VR experiences for users. Using the example for educational purposes, students can interact with historical figures in the settings that the historical figure lived in. For example, students can discuss Shakespeare's plays at the Globe Theater or learn physics from Isaac Newton under the famous apple tree. A real-life example comes from the Georgia Peanut Commission's Education Center, where deepfake and AR technology was used to animate portraits such that when visitors approach with their phones, the portrait comes to life with a short monologue (Wynn et al., 2021). An image of the frames from one of the deepfaked videos can be seen in Figure 5.1. Deepfakes in AR/VR could also make communication more engaging for users. For example, families with members spread across the globe could get together in a virtual environment where deepfakes make it look and feel like the families are together. The main issue would come from the nonconsensual use of a person's image to recreate them in a virtual environment. The more malicious uses of a person's image without their consent, such as the creation of virtual pornography using the person's image, can be damaging to that person's emotional well-being. Deepfakes in virtual reality could also be used to impersonate other people, which can be a crime and can cause financial and emotion damage to the person whose image is being reproduced and the user who ends up interacting with the fake person. Companies that include deepfakes in their AR/VR tools and products could find ways to mitigate these kinds of issues by requiring the consent of the person whose image is being used and generating models that do not lend themselves to malicious purposes.

Evaluating the potential severity of the use of deepfakes in AR/VR depends on how the deepfakes are used. For uses with positive consequences, such as for educational or communication purposes, there are many long-lasting benefits for individual users. Students will carry lessons with them through their lives

Figure 1: Frame by frame video of the image animated portrait

Figure 5.1: An image from (Wynn et al., 2021) depicting the frames from a video

because of the memorable experience they had learning from deepfaked historical figures in virtual settings. Families have the chance to connect in a way that simple phone calls or text messages cannot offer. These positive examples of detected deepfakes put severity at a 2. However, the negative consequences of the misuse of deepfakes in AR and VR, especially if they go undetected, can be severely damaging to a person's emotional and financial well-being. Damage to reputation and the possible circulation of a deepfake AR/VR model of a person without their consent are some of the possible serious issues that can arise from the use of deepfakes in AR. Companies can take steps to mitigate this issue by requiring consent and detection. The negative severity of the use of undetected deepfakes can be ranked at -3 on the scale given the potential consequences. For probability, it is very likely that companies will tap into the potential profit that can be earned through the use of deepfakes in AR/VR because the combination of those technologies would make a very believable and enjoyable product. On top of this, there is a current example of the use of deepfakes with AR for the benefit of museum visitors in Georgia. The probability can be ranked at a 1 on the scale given that deepfakes exist in AR and will see more uses within AR/VR technology and products. For detection, it is important that malicious deepfakes in AR and VR be detected, so it is likely that companies will integrate their own ways of mitigating potential malicious uses. However, it is possible that not every instance of a deepfake will be caught, especially if the deepfake AR/VR models use a combination of deepfake creation algorithms and AR/VR model algorithms. The combination of the two might not leave digital footprints to be detected. To average both sides of detection, the score will be 0.4. In total, deepfakes in AR/VR have an RPN of -1.

5.2.5 Video Games

Video games at their best create immersive and entertaining experiences for the players. The quality of the CGI to create the playable characters has improved tremendously over multiple years, and the addition of deepfake technology has the potential to make the video games even more engaging and immersive. The severity of the consequences of using deepfakes in video games would be the same as the current methods of creating characters for video games. Adding heightened realism to the graphics of a video game would enhance the experience for the players but should be done so carefully as the situations that the characters find themselves in could cause emotional or psychological harm to the player. Erring on the side of caution, the severity of the consequences of detected deepfakes in video games can be ranked at 2. Given how video games are developed, it is unlikely that the use of deepfakes in video games will go undetected. Therefore, the severity score for undetected deepfakes can be 0. Assuming that video game developers use the tool responsibly, there is little chance that there would be highly negative consequences. It is highly likely that deepfakes will be used in video games in the future. It would provide a new tool for creating immersive experiences for gamers who enjoy playing games that include human characters. On top of the immersive experience, video game developers would save time and money by having the deepfake model create the characters which they can improve with the use of CGI and special effects. To touch briefly on the job market in video game development, the inclusion of deepfakes might either eliminate or make much easier the job of character designers who might spend hours trying to create highly detailed character models for their games. Deepfakes will speed up the process for and be much cheaper in terms of cost per hour of development. The probability score for the use of deepfakes in video games is 0.8. Detection for deepfakes in video games is not necessary. The deepfakes are explicitly tools to create immersive experiences for video games, similar to how CGI is used, and it is understood that what is happening on the screen is not real. For this reason, detection has a score of 1. Overall, the use of deepfakes in video games has an RPN of 1.6.

5.2.6 Trustworthy AI

There already exists distrust in technology. A 2021 poll from the Washington Post shows that many people of all ages in the United States do not trust social media with their personal information and believe that their phones are listening to their conversations without their consent (Kelly and Guskin, 2021). This kind of distrust extends to deepfakes with how they are reported in the media, where the use of deepfakes appears in headlines for mostly negative reasons. Many of the popular media articles published tell stories about deepfaked pornography or deepfakes of political figures. The negative tone of these articles seems to promise doom, with titles ranging from “Ready or not, mass video deepfakes are coming” (Zeitchik, 2022) to “AI-Assisted Fake Porn Is Here and We’re All Fucked” (Cole, 2017) which sound ominous and overtly negative.

There are some notable exceptions, such as "Tom Cruise deepfake creator says public shouldn't be worried about 'one-click fakes'" (Vincent, 2021), but these are not the norm. Distrust also spreads because of the reporting of misinformation through the use of deepfakes (as discussed in Chapter 5.3.6). It falls on the creators of deepfakes to establish trust in the technology in the face of the large amount of negative press.

Creating trustworthy artificial intelligence means making sure that the AI models are explainable, fair, robust, safe, and transparent (Han and Choi, 2022). The Organization for Economic Co-operation and Development proposed five principles for trustworthy AI (Wickramasinghe et al., 2020), of which accountability, human-centered values and fairness, and transparency and explainability are applicable to making deepfakes a part of the trustworthy AI conversation. For accountability, the creators of the deepfake-based technologies should take responsibility for any negative consequences that the use of their deepfake causes. They can take actions to prevent the misuse of their technologies which can generate trust from users because they would know that the creators of the technology take accountability for their product. Having open communication about how deepfakes are created and where they are used introduces transparency and explainability. Products that use deepfakes can explain how they are used and how any data inputted by the user is stored, which means that users will have the option to use the product if they choose because they would understand how the technology works. Having open communication does not mean that every detail is explained, but more that an easily understandable overview and any important information is told to the user. Deepfakes are currently easy to make with phone applications and websites, with higher quality deepfakes being made in exchange for payment, so there is an element of fairness in its use already.

For human centered values, deepfake technology has the potential to introduce accessibility and diversity to its many applications. Accessibility is the ability of the deepfake to provide equitable tools for its use especially to individuals with disabilities. Using the example of education, deaf students could communicate with historical figures through sign language. Diversity is the representation of multiple genders and ethnicities in the use of deepfakes. Diversity of the generated models will be dependent on the databases that train the deepfake models. Should the deepfake models be trained on faces of one ethnicity or gender more than another, the deepfake model risks becoming biased. Ensuring that the database for training the models is diverse is necessary to create a robust model but will take time and money to create. However, by including diversity, deepfakes have the potential to bring people from around the world into the same room, giving each person an equal platform in a group setting.

Deepfakes as an example of trustworthy AI would have many benefits. Trust in the technology would lead to wider spread use, and this in turn can also create better understanding of how the technology works. Trustworthy AI is inherently a positive concept, and trust in deepfakes has potential for positive consequences should deepfake creation and use follow trustworthy AI principles. Detection becomes vital in building the

trust in deepfakes. Detection falls under the principle of transparency, where having users know that the person they might be interacting with is fake builds trust in the technology because it is not being used for deception. Given this, the severity rating for trust in detected deepfakes is 3. Deepfakes that are undetected and later found to be undetected for long periods of time could damage the trust the people have in the technology. This means that the severity of undetected deepfakes can be scored at -2. Determining the probability of increased trust in the technology is challenging. On one hand, an optimistic outlook would suggest that people will grow to trust the technology with time just as they have with other kinds of innovative technologies, such as the automobile replaced the horse and carriage. However, current attitudes towards deepfakes make it seem like building trust in the technology will take a very long time. Erring on the side of current attitudes, the probability score for trust in deepfakes is 0.4. Detection technology would assist in the building of trust in the technology, so in this case detection earns a score of 0.8. In total, trust in deepfakes has an RPN of 0.8.

5.2.7 Telehealth and Teletherapy

There is potential for deepfakes to be used in telehealth and teletherapy. During the COVID-19 pandemic, telehealth and teletherapy saw a spike in use since doctors were not able to see their patients in person due to social distancing regulations. Telehealth and teletherapy services continue being used because they provide a dimension of accessibility to health services. People who might not be able to go to the doctor's office now have the choice to meet them virtually. Deepfakes can help create more personalized experiences for the patients by creating embodied chatbots for telehealth and teletherapy services. Deepfake models can be created and used in place of text chatbots to create more comforting interactions when dealing with worrisome symptoms. Deepfake telehealth chatbots would reduce travel and wait times for patients, which in severe cases can be lifesaving, and have the potential to generate more money because they can help more patients a day without exhaustion. For teletherapy, deepfake models can be used to engage with patients should a human therapist not be there. The deepfakes, when combined with chat-based technology like ChatGPT, can be used to provide companionship for individuals seeking company or help people process their thoughts without feeling completely alone. What the medical field should be cautious of is overreliance of these tools. Patients might rely too much on deepfake doctors and therapists for diagnosis and company, which could lead to social isolation in the worst case. Deepfakes in telehealth and teletherapy should not replace doctors and therapists. Instead, the use of deepfakes should give medical professionals tools to help a larger number of people. The deepfake telehealth bots should not take the place of a proper diagnosis and should come with obviously placed notices stating that serious medical issues should be taken to human doctors.

Determining the severity of deepfakes used in telehealth and teletherapy can be done by analyzing the

uses. The use of deepfakes in telehealth and teletherapy have the potential to save lives with reduced travel and wait times for patients but can be over relied on for diagnosis and companionship. As it might be known that a deepfake is being used (the system should let the user know that the person they are talking to is virtual), the severity score of the use of detected deepfakes in telehealth can be put at a 3 because in the best cases it is saving lives. Undetected use of deepfakes in scam telehealth and teletherapy can be extremely damaging to the patients. Worst case scenarios can lead to extreme depression, anxiety, and/or suicide. Undetected use of deepfakes in telehealth and teletherapy scores a -4 on the severity scale. The probability that these deepfake telehealth bots will be used depends on how willing the medical industry is to use deepfake technology to enhance the telehealth and teletherapy experiences for patients. This means that probability can be anywhere on the scale because opinions can vary and can therefore be placed in the middle at 0.6, meaning it might happen. Finally, detection is necessary to find unsanctioned uses of deepfakes in telehealth and teletherapy because the severity of the negative consequences should not be ignored. However, clever programmers can make it difficult for a person without algorithmic detection tools to spot deepfaked telehealth services. These can be reported and taken down should they be found, but this would depend on how many people use the detection tools and choose to report the fake service. Erring on the side of caution, detection will be scored at 0.4. Overall, the use of deepfakes in telehealth and teletherapy earn an RPN of -0.72.

5.3 Threats

Deepfakes can create potential threats, many of which exploit the weaknesses of deepfakes. In this section, I will discuss some threats posed by the use of deepfakes in various fields, speculate about potential cases for the uses, and conduct an FMEA analysis of the risk posed by these threats.

5.3.1 Social Engineering

Technology can be used in different cases of social engineering. One example of this is phishing. Phishing is the use of emails by fake sources posing as credible ones to collect information from targets. Deepfakes can be used in a similar way which I will dub “puppet fraud”. Deepfakes can be made to look like credible individuals such as employees of technology companies (the “puppet”) and used to gather personal information from unwitting customers (the “fraud”). For example, a malicious entity can infiltrate a company’s help system by rerouting the chat or help system to direct people seeking assistance to a video chat with a deepfake. The deepfake can then gain the trust of the help-seeking customer by posing as an employee of the company and gathering passwords, access to the customers computer, and even credit card information under the guise of assisting the customer with their problem. Deepfakes used in social engineering can cause concern for security. For example, the deepfakes used for puppet fraud have to be believable to the human eye without

the use of detection software, meaning that unwitting technology users might be fooled into giving away their sensitive information to the deepfake. In cases like these, implementing dual factor authentication, such as Duo or simple authentication codes texted to phones, might help add a level of security. However, these can be circumvented by the deepfake model if the technology user that the deepfake is talking to is convinced by the deepfake to confirm the authentication.

The loss of security through puppet fraud can lead to extreme financial harm that can only be undone through the law. The financial harm can also lead to emotional harm for the person who was targeted by the deepfake. Companies might also suffer from loss of reputation, extreme damage to their security, and potential financial loss as a result of the discovery of the deepfake in their system. The severity of the damages posed by social engineering via deepfakes should the deepfakes be left undetected places their severity score at -4. The severity of the deepfakes should they be detected depends on how quickly they are caught and how many people interact with the deepfake before it is caught. Assuming the case (to give contrast to the worst-case undetected severity) where the deepfake is caught in a timely manner but was still interacted with, the severity of the deepfake having been detected would be -1. It is likely that this use of deepfakes will get caught because effected customers will start reporting the damages as soon as they see. This will prompt the companies to check their systems and shore up their defenses. However, if scammers chose to pose as original or little-known companies (not established and well-known companies) and create the fake resources for users to contact, users might still fall for the scam. For this reason, the detection score for social engineering like puppet fraud can be ranked at 0.4. As the technology stands right now, it would be very challenging but not impossible to create such realistic puppets. As Chris Ume describes, creating realistic deepfakes that completely fool the viewer takes a combination of deep learning, special effects, and a lot of time (Vincent, 2021). There is also the risk of getting caught very quickly if companies use deepfake detection software in their code to ensure that only reputable partners and their own employees have access to their code. These two factors combined make it unlikely but not impossible that we might see cases of puppet fraud in the future, putting its probability score at 0.6. Overall, social engineering such as puppet fraud has an RPN of -1.68.

5.3.2 Consent for Image Use

Ideally, written consent is given for someone's image to be used for various purposes, especially in marketing. In the case of deepfakes, the use of someone's image without their consent becomes a tricky topic because of the various applications of deepfakes. In cases of parody and fair use, creators of the deepfakes might not need permission for someone's image to be used. For research, the use of a person's image can help train a model. In cases such as the use of deepfakes in defamation, legal action can be taken against the

creators because there are various state laws that protect a person's image (these are discussed in detail in Chapter 5.3.3). Nonconsensual use of a person's image can lend itself for especially malicious purposes such as revenge pornography, which has severe emotional, societal, and potentially financial repercussions for the target of the deepfake.

Celebrities are generally the targets of deepfakes that use their image without their explicit consent. Deepfake algorithms, like many machine learning algorithms, need to train on a large number of training images, which there are many available of celebrities in entertainment, politics, and social media. There are examples both in popular media and in research of the image of celebrities being used to create deepfakes. One recent example is of the TikTok account DeepTomCruise and its use of deepfakes to create humorous videos with a stand-in actor using the face of actor Tom Cruise (as pictured in Figure 5.2). In an interview with The Verge, DeepTomCruise creator and special effects artist Chris Ume explains that the process to create such realistic deepfakes takes weeks for each clip (Vincent, 2021), most of which are under one minute long. Ume creates these short videos using DeepFaceLab's open-source algorithm and video editing tools to "make sure you don't see any of the glitches" (Vincent, 2021). The purpose of DeepTomCruise was to create awareness of deepfake technology while having fun and showing off Ume's technical skills (Vincent, 2021), even appearing on the television show America's Got Talent in August of 2022 to excited responses from the show's judges and horrified responses from critics (Zeitchik, 2022). Other examples of celebrity images being used for entertaining or parody deepfakes can be found online.

Want to see a magic trick? Tom Cruise impersonator Miles Fisher (left) and the deepfake Tom Cruise created by Chris Ume (right). Image: Chris Ume

Figure 5.2: Stand-in actor Miles Fisher (left) and deepfake Tom Cruise (right) as pictured in (Vincent, 2021)

Celebrity images are used in research because there are many images that are easily available online. (Li et al., 2020) introduces Celeb-DF, a database of 5,639 deepfake videos of celebrities meant for "the

development and evaluation of DeepFake detection algorithms" which has been used in research such as (Vamsi et al., 2022), (Ciftci et al., 2020), and (Al-Dhabi and Zhang, 2021). The deepfake videos in the Celeb-DF dataset were created using face-swapping on pairs of the 59 subjects in 590 real publicly available videos on YouTube (these real videos are also part of the dataset) (Li et al., 2020). Face2Face, a landmark paper from 2016 which proposed a novel approach for expression puppetry, used celebrity images to test their method. The researchers used video clips from YouTube which "show different subjects in different scenes filmed from varying camera angles…", and the figure showing the results depicts former Presidents George W. Bush, Donald Trump, and actor Daniel Craig (Thies et al., 2016). As the research is not monetized or used for commercial purposes, researchers are within their rights to use celebrities' images for training and testing their models.

The severity of the uses of deepfakes in the context of consent depend on if consent is obtained or not. Obtaining consent for the use of a person's image does not have many, if any, negative consequences outside of the person refusing to give permission. The act of obtaining consent to use someone's image means that it is known that a deepfake is getting created and are hence "detected". It is unlikely for deepfakes that have the consent of the person in the video to be undetected. For these reasons, the risk evaluation falls onto the risk of not obtaining consent for the use of someone's likeness. The lack of consent in the current uses of deepfakes generates distrust from the public, especially when the images are used to train models that depict the target saying or doing something that potentially has negative consequences. Should these go undetected, the severity of the consequences could be ranked on the severity scale at -3. These deepfakes can be detected and marked as fake, but would still have potential emotional, societal, and financial harm for the target depicted in the video. Therefore, the score for detected deepfakes that do not have consent would be -1. The probability that the lack of consent for the use of a person's image in a deepfake will continue is high because of the kinds of deepfakes being made. For parody or satire (which many popular deepfakes are), there might not be a need to obtain consent because the person themself is not being attached to a marketed message, and it is usually clear that the parody or satire is not actually the person themselves. For research, the data is used for training, not marketing, and many databases already exist that have the images available to use. For these reasons, consent in the use of deepfakes has a probability of 0.8. For deepfakes created without the consent of the person pictured, it would be important that the person being depicted knows that their image is being used especially if it might be used for a malicious purpose. Detection becomes vital in this case, so it would be very likely that deepfake detection tools would be used to monitor various online locations for uses of deepfakes that might contain the image of someone who has not consented to the use of their image. However, it would be more likely that famous faces would be caught with ease, while non-famous face (which is a larger population) might not be caught. This would make detection 0.2 on the scale.

Overall, the RPN of this threat is -2.08.

5.3.3 Deepfakes and the Law

There are two notable connections between deepfakes and the law. The first is how the law might try to govern the creation and use of deepfakes. The second is how deepfakes might be used in a court of law.

It is difficult to regulate technology. The term "regulation" in the case of deepfakes means the mitigation of the creation and distribution of the technology itself. At the federal level, it would take Congress members to first understand what deepfakes are and how they are used, then agree on a set of limits. Congress members would have to not encroach on fair use, which protects the parodies and satire that some deepfakes currently are. Federal regulations regarding deepfakes could take inspiration from current state regulations. At the state level, some states have laws that address deepfakes either directly or indirectly. For example, California Assembly Bill 730 from 2019 addresses the use of deepfakes during elections. The bill "would prohibit a person, committee, or other entity, within 60 days of an election at which a candidate for elective office will appear on the ballot, from distributing with actual malice materially deceptive audio or visual media of the candidate with the intent to injure the candidate's reputation or to deceive a voter into voting for or against the candidate, unless the media includes a disclosure stating that the media has been manipulated" (Berman and Grayson, 2019). Another example is Texas Penal Code 33.07, which states that it is a felony if a person uses the name or persona of another person "with the intent to harm, defraud, intimidate, or threaten any person" without the consent of the person whose image is being used (82nd Legislature of Texas, 2021).

There is also the question of copyright law and how it might affect the creation and output of deepfakes. U.S. Copyright Law generally protects "original works of authorship fixed in any tangible medium of expression...from which they can be perceived, reproduced, or otherwise communicated either directly or with the aid of a machine or device", which includes pictorial, motion picture, and other audiovisual works (Office, 2022). Depending on what kind of deepfake has been made, the video could potentially fall into two categories: original work or derivative work. Deepfakes that are completely original creations, meaning that the video does not rely on the manipulation of a previously existing piece of media, might fall under copyright protection because it can be considered an original audiovisual work. Deepfakes could be considered a form of "derivative work", defined by U.S. Copyright Law as "a work based upon one or more preexisting works"(Office, 2022), as some deepfakes like face swapping are based on a previously existing picture or video. These could still be considered protected under copyright law as long as the deepfake creator does not try to claim credit for the original source material that the deepfake is based on. Should the deepfake be a derivative work, the creator of the deepfake would have to consider the copyright of the original material in order to avoid potential conflicts. The actual effects of copyright law on deepfakes depend on how courts

of law determine copyright laws to apply. Slowly but surely, governments at various levels are attempting to mitigate the creation and distribution of harmful deepfakes without stopping the development of beneficial or otherwise neutral deepfakes.

As deepfakes become more realistic, there is a chance that they might be used to falsify evidence in cases of law. For example, a guilty defendant could submit a deepfaked video of themselves in a location different from the crime scene to establish an alibi so that they are found innocent. A vindictive plaintiff could do the opposite, creating a deepfake of the defendant committing the crime. Should the deepfaked evidence not be detected, the guilty party could go free. Having undetected deepfaked evidence in cases of law would have serious consequences for an individual, making it a -4 on the severity scale. There are two factors that can counter the use of deepfakes in the law. First are the current deepfake detection methods available. The deepfake detection research discussed earlier shows that most methods are highly accurate, so when applied to videos applied in cases of law, a deepfaked piece of evidence is likely to get caught. The second factor is the law itself. Knowingly submitting false evidence in court is a felony that can lead to prison time of up to five years, a fine, or both (DOJ, 2020). Adding on, should the defendant be found to have submitted falsified evidence, it would cast further doubt on their innocence. Detected deepfake evidence in law would almost definitely have large negative consequences for the team that submitted it and potentially positive outcomes for the opposing team. Detected deepfakes in law earn a -1 on the severity scale to strike a balance between the negative outcomes for the deepfake user and the positive outcomes for their opposition. The consequences of the submitting deepfaked evidence should deter people from doing it, making the probability of this occurring in the future low. Deepfaked evidence would the probability scale at a 0.2 because while it would not be likely that it would happen, people still might try to pass off deepfaked evidence as real. Courts should look into using deepfake detection technology to ensure that any image or video evidence they receive is real and credible. Without detection, the severity of the consequences become worse for the parties affected so it is very likely that deepfake detection technology will be used by the courts to verify evidence. With the reported accuracy of deepfake detection methods and the likelihood that detection methods will be used, this earns deepfakes in law a score of 0.8 for detection. Overall, deepfakes and the law has an RPN of -0.32.

5.3.4 Online Harassment

Online harassment is not a novel concept. It can be defined as a wide range of behaviors enabled by various pieces of technology used to target a user or users (Blackwell et al., 2017). In 2017, a Pew survey reported that 66% of adult internet users had witnessed online harassment, and 41% of users had been harassed online (Blackwell et al., 2017). Deepfakes can potentially add to the severity of online harassment. Deepfakes can be created without the consent of an individual and used for purposes of blackmail and impersonation.

For example, realistic videos of the person doing or saying something that they did not actually say or do can be created and used as leverage against the target of the video. Examples of this already exist in the form of deepfake pornography, which can be considered another form of online harassment. This problem can especially affect celebrities, whose numerous images available easily online lend themselves easily to deepfake models for training. These deepfakes would be harder to disprove than technology that leaves paper trails such as social media. In addition, deepfakes used in online harassment such as blackmail would result in severe emotional and financial damage for the target. Impersonation of the target for malicious purposes could also lead to personal and societal damage if the target is not made immediately aware of how their image is being used. This could lead to isolation and depression in the victim, which themselves could have serious consequences. Victims of current forms of online harassment experiencing disruption from everyday life and have to spend time reporting the harassment, which takes away from their personal responsibilities, work obligations, and even sleep (Blackwell et al., 2017). This would worsen with the use of deepfakes because now the victim is aware of their image being used maliciously. As mentioned previously, there are laws that protect a person's image. However, it would take first detecting the misuse of the deepfake then locating the creator in order for the law to apply.

There are extreme severe consequences for the use of deepfakes for online harassment. Immediate consequences would be emotional, financial, and societal harm for the victim. These can lead to further psychological damage that would either take years to repair or simply be irreparable. Due to the severity of the consequences, online harassment using undetected deepfakes is a -4 on the severity scale. Should the deepfakes be detected, the consequences depend on the amount of time that has passed. To provide a contrast to undetected, detected deepfakes have a score of -1 where it is assumed that they were caught in a timely manner, but damage was still done. The probability that online harassment with deepfakes continues is very high because the technology exists, and examples have already been created in the past. This places the probability score at a 0.8. Finally, in order to mitigate the consequences of deepfake online harassment, it is vital that detection methods are spread to any possible platform that this kind of harassment can occur. It is likely that deepfakes of popular figures will get caught with these detection methods as many of them are trained on these popular faces. However, this leaves people in the general public more vulnerable. The targeted person who is depicted in the deepfake would most likely say that the video is deepfaked, meaning it has been caught, but it might take additional algorithmic tools to have others believe the targeted person. This places the detection score at 0.4, making the overall RPN of deepfakes in online harassment -2.24.

5.3.5 Deepfake Pornography

GAN technology has been used to create deepfakes of popular actors in pornographic material. Generally, celebrities and other popular figures are targets of this type of deepfake because there are a lot of images and videos of them available online. The first well-known instance of this was in 2017, when the face of actress Gal Gadot was used in an existing pornographic video and posted to Reddit by the user "deepfakes" who claimed to be the creator of the video (DHS, 2019). A more recent example is from February of 2023, when Twitch streamer "Sweet Anita" found out that her likeness was used in creating deepfaked pornography (O'Sullivan, 2023). The nonconsensual use of a person's likeness in sensitive material can be emotionally distressing for the target, who can take legal action. Lawmakers in California have created laws to "try to counter the potential for deepfakes to be used in an election campaign and in nonconsensual pornography" (O'Sullivan, 2023). California Assembly Bill 602 from 2019 provides a right of action against a person who "Creates and intentionally discloses sexually explicit material and the person knows or reasonably should have known the depicted individual in that material did not consent to its creation or disclosure" with plaintiff being able to recover money, economic and non-economic damages, punitive damages and more (Berman and Leyva, 2019). To briefly touch upon the job market, deepfakes have a very similar effect here as they do in entertainment. Deepfakes might be used for face swapping actors faces in pornographic material, which might still leave the need for a stand-in actor.

There are a variety of consequences for the creation and circulation of deepfake pornography. It can damage the reputation of the person in the video which, for celebrities who rely on marketing their image and members of the public who value their privacy, is a concerning problem. While there may be laws which attempt to protect people from having their likeness used without their consent, it does not guarantee that people will stop making these kinds of videos. Another problem is that once the deepfakes are online, they are impossible to remove completely. Copies likely exist somewhere online even after the original is taken down. The creation and spread of these deepfakes, especially as they improve and become harder to detect as fake with the human eye alone, can be extremely emotionally damaging for the targets. Deepfake pornography can be made of anybody should the algorithm be given enough training material with the target's image which means that the risk could apply to anyone. Furthermore, there is little protection for the people whose images appear in deepfakes. Governments are the verified sources that can confirm or deny the information in a deepfake of a politician, but individual people do not have the same kind of protection. Should the deepfake go undetected, the damage to the livelihood and reputation of the people whose likeness is used in the deepfake, the concern of having deepfake pornography constantly spread, and the negative emotional damage caused by the creation and circulation of the deepfake puts this use at a -4 on the severity scale.

Detection of the deepfake would probably mean that the original deepfake gets taken down, but that does not stop the copies from spreading. On top of that, the amount of time that it would take for the average person to recover from emotional or interpersonal damage might be very long depending on where they are and the community around them. Detection might be helpful in mitigating some of the severity, but there is a chance that damage recovery might take time, so the severity score for detected deepfake pornography can be scored at -2. The probability that deepfake pornography will keep being made is high and will remain high until laws make the creation and distribution of deepfake pornography a felony. As of 2022, 96% of deepfakes came from pornographic material (Malik et al., 2022). This puts deepfake pornography at a 1 currently on the probability scale. Detection is ideal for deepfaked pornography, but it might be a long time before it is implemented. With so many deepfakes being pornographic materials, it might be time consuming to figure out if the deepfakes are of real people or not. It would cost companies money and time to implement the detection tools on their websites, at which point it becomes a question of if the companies are willing to spend the time and money. Therefore, the score for detection could be 0.4. Overall, deepfake pornography has an RPN of -3.2.

5.3.6 Misinformation

It is not uncommon to hear about the use of technology to spread the wrong information to the public. With the introduction of deepfakes, people now have another method to spread information that serves their own interests that might not align with the truth. One example of this is the use of lifelike deepfakes by Spamouflage, a pro-Chinese political spam operation, to create videos promoting the interests of the Chinese Communist Party under the guise of a media company called “Wolf News” (Graphika, 2023). According to a report by the intelligence company Graphika, “Wolf News” uses the deepfake models from the British AI video company Synthesia to create videos that look like real news broadcasts (Graphika, 2023). The models created and marketed by Synthesia are lifelike at first glance but fall into the Uncanny Valley because of the irregular mouth and eye movement that can be seen when the models move. Should the viewer not be watching carefully and engaging in critical thinking, the deepfake video could be perceived by the viewer as fact. Misinformation spread through deepfakes can also contribute to confirmation bias. Confirmation bias is defined as a kind of bias where a person seeks out, interprets, or prioritizes information that bolsters their own set of beliefs, opinions, or hypotheses (Rafezi et al., 2019), (Suzuki and Yamamoto, 2021). There is a chance that people who are exposed to deepfakes that bolster their beliefs might not critically examine the deepfake for signs of falsehood. An occurrence of this is reported in the experiment from (Suzuki and Yamamoto, 2021), where the results showed that the participants who had confirmation bias could not make effective use of their health literacy (defined in the paper as “the skill to search for trustworthy health information on the

web") while conducting web searches about health information. Deepfakes that state views that are contrary to a person's belief and are proven to be fake might cause the person to strengthen their own belief even if that belief is factually incorrect. This can also create a general distrust in the technology which can be detrimental to its continued use.

The potential misinformation that can be spread by realistic deepfakes has overtly negative consequences. The spread of misinformation can be damaging to a person's reputation, can result in dangerous actions based on incorrect information, and can be used to influence world altering events. There are no benefits of the spread of misinformation through deepfakes and the negative consequences have the potential to be catastrophic not only for individuals but also for larger groups especially if the deepfake goes undetected. For this reason, undetected deepfakes that spread misinformation can be scored at -4 on the severity scale. If the deepfake is detected, it still has the risk of spreading the information before it is caught. The deepfake can also be countered by explanations using factual information and can be reported to the website it is posed on. If it is caught before it is uploaded, then the consequences are reduced greatly. To average these two factors, detected deepfakes earn a score of -1 on the detection scale. There is a high chance that deepfakes will continue to be used to spread misinformation or be used to serve the specific purposes of the individual or group making and distributing them. There is a current example of this with Spamouflage and "Wolf News". This means that the probability score for deepfakes and misinformation is 0.8. The severity of misinformation spread by deepfakes is high and therefore warrants prompt detection. The consequences of leaving misinformation spreading deepfakes undetected could lead to catastrophic consequences for individuals and groups that potentially cannot be fixed. If online sources employ deepfake detection algorithms to user inputs and users learn what visual cues indicate a video is a deepfake, then it is more likely that the deepfake will get caught. Otherwise, it will take experts and group consensus to figure out videos are fake. Taking a more cautious approach, the detection score can be placed at 0.4, making the RPN -2.24.

5.3.7 Deepfakes in Politics

There have been notable deepfake videos made of politicians. Some of these are made for humorous or informational purposes. One famous example from 2018 comes from BuzzFeed. In a video titled, "You Won't Believe What Obama Says In This Video!", a deepfake of former United States President Barack Obama is depicted cautioning about the spread of deepfake technology (BuzzFeedVideo, 2018). It is revealed about halfway through the video that it is not President Obama, but filmmaker Jordan Peele who is speaking to the audience watching the video (a screenshot of this can be seen in Figure 5.3). Peele's Obama cautions "Moving forward, we need to be more vigilant with what we trust from the Internet" (BuzzFeedVideo, 2018). Other deepfake videos of politicians have the potential to cause great harm in dangerous situations. In March of

2022, a deepfake of Ukrainian President Volodymyr Zelenskyy calling on Ukrainian soldiers to stop fighting spread on social media and Ukrainian news before being identified as fake and taken down (Allyn, 2022). This instance of the spread of deepfakes causes distrust in political media: Professor Hany Farid from the University of California, Berkeley's School of Information explains, "The next time the president goes on television, some people might think, 'Wait a minute — is this real?'" (Allyn, 2022). In response to the circulation of deepfaked videos, a spokesperson from Twitter explained that the platform allows deepfakes in instances where it is being shown to be fake but will be taken down if it is used for deception (Allyn, 2022).

Figure 5.3: A screenshot from (BuzzFeedVideo, 2018)

Deepfakes used in politics merit the use of detection tools as they have the potential to cause severe harm if they are not caught. There would have been the chance of Ukrainian soldiers laying down arms and surrendering had the deepfake video of President Zelenskyy not been marked as a deepfake, potentially losing the war for Ukraine and causing further irreparable damage to individual lives. Consumers of news media would have to stay constantly vigilant and check every piece of news they see and hear to confirm the facts, which defeats the purpose of having credible news sources. Politicians also risk losing their credibility with the public and the media should deepfakes not be detected. The potential harms deepfakes pose in politics could place their severity high on the severity scale as the consequences of deepfakes can be catastrophic for people in countries at war and severely damaging in elections. Erring on the side of caution and judging use of deepfakes in politics based on potential severity of the consequences, the severity of undetected deepfakes in politics can be placed at -4 on the severity scale. On the other hand, deepfakes of politicians make the news because they have been detected. The extreme negative consequences of these deepfakes are generally avoided because of detection and fact-checking. The deepfake of President Zelenskyy did not cause Ukraine to surrender because it was detected and marked as fake. With current deepfake detection technology as it is available and the quick response of government bodies, the consequences posed by the use of deepfakes are

mitigated. Keeping in mind that the consequences have the potential to be severe but historically have not because of detection methods in place, the severity of detected deepfakes in politics can be scored at -1. As media sees an increase in the number of deepfakes, it is very likely that deepfakes of politicians will continue to be used and identified. This places deepfakes in politics at a 0.8 on the probability scale. Detection will be vital for deepfakes in politics in the future. The threat of the spread of misinformation through deepfakes politicians is enough to warrant detection algorithms being used in places where the videos of deepfaked politicians can spread. On top of this, because politicians are well known figures, it would be easy to find and check videos of the politicians for deepfakes though one should acknowledge that a few might slip through detection methods. This earns a detection score of 0.8. Overall, deepfakes in politics has an RPN of -1.28.

CHAPTER 6

Results

The RPNs of each of the opportunities and threats have been listed in the tables below. The tables are categorized according to if the item was an opportunity or threat. The averages of the severity, probability, detection, and RPN scores are at the bottom of each column. There are 5 items that pose little to no risk and 9 items that pose a risk. Deepfake pornography poses the most severe risk with an RPN of -3.2. The least risk is posed by the opportunity created by entertainment, which has an RPN of 1.76.

Opportunities	**Severity (Detected)**	**Severity (Undetected)**	**Probability**	**Detection**	**RPN**
Fashion	2	-2	0.8	0.6	0.4
Entertainment	3	-1	0.8	0.8	1.76
Education	3	-2	0.6	0.6	0.6
AR/VR	2	-3	1	0.4	-1
Video Games	2	0	0.8	1	1.6
Trustworthy AI	3	-2	0.4	0.8	0.8
Telehealth	3	-4	0.6	0.4	-0.72
Average	2.57	-2.00	0.71	0.66	0.49

Table 6.1: Breakdown of RPNs for the opportunities of deepfakes

The scores for opportunities are listed in Table 6.1. The severity score for detected deepfakes is positive which means that there are some benefits provided by the opportunities for the use of deepfakes. The severity score for undetected deepfakes is negative as the consequences of leaving the deepfakes undetected are not negligible and can be severe. The average probability that deepfakes will be used in the opportunities listed is high, meaning that there is a very likely chance that deepfakes will be used. The detection score indicates that if deepfake detection tools will be used within these opportunities in the future, the deepfakes will be easier than not to catch, though some might slip past detection methods. Finally, the average RPN is positive, indicating that there is little to no risk posed by the opportunities created by deepfakes. However, compared to the potential maximum positive RPN score, this is a small value, which means that the uses created by the opportunities of deepfakes have very few overall benefits for the end users. It is worth noting that some of the RPNs are negative, meaning that the negative consequences, especially if the deepfake goes undetected, outweigh the benefits.

The scores for threats of deepfakes are listed in Table 6.2. The severity score of undetected and detected deepfakes is negative because the threats posed have damaging emotional, financial, and societal consequences for the targets. Undetected deepfakes have a much larger negative score than detected deepfakes

Threats	Severity (Detected)	Severity (Undetected)	Probability	Detection	RPN
Social Engineering	-1	-4	0.6	0.4	-1.68
Consent	-1	-3	0.8	0.2	-2.08
Law	-1	-4	0.2	0.8	-0.32
Online Harassment	-1	-4	0.8	0.4	-2.24
Pornography	-2	-4	1	0.4	-3.2
Misinformation	-1	-4	0.8	0.4	-2.24
Politics	-1	-4	0.8	0.8	-1.28
Average	-1.14	-3.86	0.71	0.49	-1.86

Table 6.2: Breakdown of RPNs for the threats of deepfakes

because undetected deepfakes have more severely damaging consequences. There is a likely chance that these threats will be seen in the future as indicated by the probability score. The detection score indicates that uses of deepfakes that pose threats may or may not be caught, likely due to the fact that most of the uses listed range from a little difficult to difficult to detect. The average RPN of the threats of deepfakes is a negative value, which indicates that the threats posed by deepfakes are concerning and should be mitigated before the technology is available for public use.

SWOT	Severity (Detected)	Severity (Undetected)	Probability	Detection	Reported RPN	Calculated RPN
Opportunities	2.57	-2.00	0.71	0.66	0.49	0.72
Threats	-1.14	-3.86	0.71	0.49	-1.86	-1.76
Average	0.72	-2.93	0.71	0.58	-0.69	-0.52

Table 6.3: Summary of Average Scores

Table 6.3 shows the average scores for the severity, probability, detection, and RPN of the opportunities and threats. There are two RPNs listed: the reported RPN which is copied from the previous tables, and the calculated RPN which is calculated with the averaged severity, probability, and detection score for the opportunities and threats.

Overall, the average severity posed by the opportunities and threats of undetected deepfakes is very severe according to the severity scale. This indicates that the consequences for undetected deepfake uses pose a negative emotional, financial, or societal harm that is not easily reparable with time, as according to Table 5.3. For opportunities, the score indicates that there is damage that can be done, but said damage is reparable with time. For threats, the score indicates that the guaranteed damage done might not be so easily reparable. The average severity of detected deepfakes is a low positive value, which means that deepfakes, when detected and known, have a few benefits that can leave positive emotional, financial, and social outcomes. Opportunities,

in their positive uses, would be known to use deepfakes or not have severe consequences should the positive uses be detected, and therefore the score reflects that there are some benefits. For threats, damage could potentially be done before the detection, but the act of detection could mitigate some of the consequences, which is indicated by the score. The probability that deepfakes will be seen in the future is very likely according to the probability scale for opportunities, threats, and their average. This indicates that deepfakes, in their many uses, will be seen in the future at some point. The average score for detection of deepfakes, according to Table 5.3, is between a little easy and a little difficult. Some deepfakes might be caught while others might not, which can be seen reflected in the scores for opportunities (which leans towards more being detected) and threats (which leans towards less).

Both the average reported RPN and average calculated RPN indicate that the various opportunities and threats of deepfakes pose a risk. According to the severity scale, the RPNs indicate that there is mild emotional, financial, or societal damage that the deepfakes can cause, but this damage is easily reparable. The reported RPN indicates that the overall threat is a very small amount greater than what is indicated by the calculated RPN. Dissecting the scores, threats of deepfakes have the most severe risk if they are used because their consequences are extremely personally, financially, and societally harmful for an individual, the probability that they will be seen in the future is likely, and there are major negative consequences should they be left undetected. Both the calculated and reported RPNs for the threats of deepfakes are negative, and both indicate that there is a non-negligible risk. The opportunities have the least risk, offering more benefits in their very likely use in the future and not having many negative consequences that require the aggressive use of detection methods. The calculated and reported RPNs for opportunities are small positive values, indicating that there is very little benefit offered by the opportunities.

CHAPTER 7

Conclusion

This thesis discussed just a few of the potentially many strengths, weaknesses, opportunities, and threats of deepfakes. Future studies would have the benefit of seeing the developments in the field of deepfakes and would be able to add to the list presented in this thesis. Another improvement to the method used here is adapting the severity, probability, and detection scales to include factors that were left out such as time needed and prevalence. Adding these two factors, along with others that were excluded, would provide a more thorough analysis of the risk posed by deepfakes. There is also space for Human Computer Interaction (HCI) analysis. As deepfakes are a relatively new technology, the speculation of their uses and effects can fall under the umbrella of design fiction. There is also potential for human studies of user interaction with deepfakes that can give more concrete values for framework analysis. For example, a survey of user interactions and opinions about deepfakes could change the severity, probability, and detection scores listed in the FMEA analysis in this thesis.

The evolution of any kind of technology is uncertain. Deepfakes are viewed in a more negative light, which is understandable given the risk that they potentially pose. The technology itself is an application of GANs and is neutral in its existence. The use that deepfake creators are putting deepfake technology to generates many negative consequences that creates risk, causing fear and distrust in the technology. This negative approach to the concept of deepfakes prevents people from seeing the positive uses such technology can have in the future. However, any use of deepfakes should be approached carefully because of the risk posed by the uses. In this thesis, I described what deepfakes are, how they are created and detected, and previous risk analysis work to give a brief background of the technology. I then conducted a SWOT analysis of deepfakes to give a list of the strengths, weaknesses, opportunities, and threats of deepfakes. Using an FMEA style analysis, I analyzed the risk posed by the uses listed under opportunities and threats of deepfakes, giving each a numeric score for the severity of the consequences, the probability that the item will happen, and the potential consequences should the item go undetected. The final analysis showed that deepfakes do pose a risk, meaning that it is important to include risk mitigation and evaluation of the potential uses of any future deepfake technology.

Future development for deepfakes should be done with the potential risks and human centered values in mind. To address the risks of misuse of the technology during the creation process, deepfake creators can incorporate mitigation methods such as requiring the consent of the person whose image is used, checking databases for the kind of content they have, and having detection methods available for public use. Mitigation

can also come from the law. There are previously mentioned state laws that can be applied to deepfakes, but there is potential for laws to be created that specifically address deepfakes. For example, a law (either at the federal or state level) could be created that would require deepfakes creators to label their videos as deepfakes if they fall under topics that have the potential to spread misinformation such as politics. These kinds of mitigation methods would generate more trust in the deepfake tools because they show that the people involved in creating deepfakes have the end users' interests in mind. Designing deepfake technology to be open about the kind of data it uses and how it is trained should also be an important part of the development process. Finally, there should be a shift in the way deepfakes are discussed in media. The negative way that deepfakes are discussed, while understandable given the risk they pose, hinders the discussion of the beneficial uses. Creation and increased discussion of the positive uses and the strengths of deepfakes would give more arguments in favor of their use. While concern about technology is good and keeps development in check, it is more productive to view the technology with optimism for what kinds of applications it might have in the future.

References

82nd Legislature of Texas (2021). Penal code sec. 33.07. online impersonation.

A. Safonov, M., S. Usov, S., and V. Arkhipov, S. (2021). E-learning application effectiveness in higher education. general research based on swot analysis. In *Proceedings of the 5th International Conference on Education and Multimedia Technology*, ICEMT '21, page 207–212, New York, NY, USA. Association for Computing Machinery.

Ahmed, M. I. and Kumar, R. (2022). A distributed generation framework using swot analysis. In *2022 2nd International Conference on Emerging Frontiers in Electrical and Electronic Technologies (ICEFEET)*, pages 1–5.

Al-Dhabi, Y. and Zhang, S. (2021). Deepfake video detection by combining convolutional neural network (cnn) and recurrent neural network (rnn). In *2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE)*, pages 236–241.

Ali, A., Khan Ghouri, K. F., Naseem, H., Soomro, T. R., Mansoor, W., and Momani, A. M. (2022). Battle of deep fakes: Artificial intelligence set to become a major threat to the individual and national security. In *2022 International Conference on Cyber Resilience (ICCR)*, pages 1–5.

Allyn, B. (2022). Deepfake video of zelenskyy could be 'tip of the iceberg' in info war, experts warn. *NPR*.

Berman, M. and Grayson, T. (2019). Ab-730 elections: deceptive audio or visual media.

Berman, M. and Leyva, C. (2019). Ab-602 depiction of individual using digital or electronic technology: sexually explicit material: cause of action.

Blackwell, L., Dimond, J., Schoenebeck, S., and Lampe, C. (2017). Classification and its consequences for online harassment: Design insights from heartmob. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW).

BuzzFeedVideo (2018). You won't believe what obama says in this video!

Chadha, A., Kumar, V., Kashyap, S., and Gupta, M. (2021). Deepfake: An overview. In Singh, P. K., Wierzchoń, S. T., Tanwar, S., Ganzha, M., and Rodrigues, J. J. P. C., editors, *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security*, pages 557–566, Singapore. Springer Singapore.

Chrysler Corporation, Ford Motor Company, G. M. C. (1995). Potential failure mode and effects analysis (fmea) reference manual.

Ciftci, U. A., Demir, I., and Yin, L. (2020). Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

CMS (2021). Guidance for performing failure mode and effects analysis with performance improvement projects.

Cole, S. (2017). Ai-assisted fake porn is here and we're all fucked. *Motherboard*.

DHS (2019). *Increasing Threat of Deepfake Identities*. Department of Homeland Security.

DOJ (2020). *902. 1996 amendments to 18 U.S.C. § 1001*. The United States Department of Justice.

Fan, Y., Xie, M., Wu, P., and Yang, G. (2022). Real-time deepfake system for live streaming. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, ICMR '22, page 202–205, New York, NY, USA. Association for Computing Machinery.

Gamage, D., Ghasiya, P., Bonagiri, V., Whiting, M. E., and Sasahara, K. (2022). Are deepfakes concerning? analyzing conversations of deepfakes on reddit and exploring societal implications. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Gandhi, A. and Jain, S. (2020). Adversarial perturbations fool deepfake detectors. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Goddard, P. (1993). Validating the safety of embedded real-time control systems using fmea. In *Annual Reliability and Maintainability Symposium 1993 Proceedings*, pages 227–230.

Goddard, P. (2000). Software fmea techniques. In *Annual Reliability and Maintainability Symposium. 2000 Proceedings. International Symposium on Product Quality and Integrity (Cat. No.00CH37055)*, pages 118–123.

Google (2023). Google trends.

Gottfried, J. (2019). About three-quarters of americans favor steps to restrict altered videos and images. *Pew Research Center*.

Graphika (2023). *Graphika Report: Deepfake It Till You Make It*.

Groh, M., Epstein, Z., Firestone, C., and Picard, R. (2021). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1).

Güera, D. and Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6.

Han, S.-H. and Choi, H.-J. (2022). Checklist for validating trustworthy ai. In *2022 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 391–394.

Hanspal, A. (2021). Here's why robots are actually going to increase human employment. *World Economic Forum*.

Hariharan, B., S, K., S, I. P., Nalina, E., N. R, W. B., and Senthil Prakash, P. N. (2022). Hybrid deep convolutional generative adversarial networks (dcgans) and style generative adversarial network (stylegans) algorithms to improve image quality. In *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 1182–1186.

Hertzum, M., Clemmensen, T., Campos, P. F., Barricelli, B. R., Hansen, C. E. D., Herbæk, L. K., Abdelnour-Nocera, J., Lopes, A. G., and Saadati, P. (2023). A swot analysis of pilot implementation. *Interactions*, 30(1):36–41.

Hunt, J. (2022). Mandalorian's luke skywalker without cgi: Mark hamill, deep fake & deaging. *Screen Rant*.

IMDb (2023). Roman holiday.

Jiwtode, M., Asati, A., Kamble, S., and Damahe, L. (2022). Deepfake video detection using neural networks. In *2022 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS)*, pages 1–5.

Kelly, H. and Guskin, E. (2021). Americans widely distrust facebook, tiktok and instagram with their data, poll finds. *The Washington Post*.

Khalil, H. A. and Maged, S. A. (2021). Deepfakes creation and detection using deep learning. In *2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pages 1–4.

Kharchenko, N. L., Izmailov, A. Z., Varfolomeeva, N. S., Bagdasarova, I. Y., Lutsenko, N. S., and Pozhidaeva, E. I. (2022). Swot analysis method application in assessing the effectiveness of moodle platform. In *2022 13th International Conference on E-Education, E-Business, E-Management, and E-Learning (IC4E)*, IC4E 2022, page 253–257, New York, NY, USA. Association for Computing Machinery.

Kim, H. H. (2014). Sw fmea for iso-26262 software development. In *2014 21st Asia-Pacific Software Engineering Conference*, volume 2, pages 19–22.

Koopman, M., Macarulla Rodriguez, A., and Geradts, Z. (2018). Detection of deepfake video manipulation. In *Proceedings of the 20th Irish Machine Vision and Image Processing conference (IMVIP)*, pages 133–136.

Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3204–3213.

Lyu, S. (2020). Deepfake detection: Current challenges and next steps. In *2020 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 1–6.

Malik, A., Kuribayashi, M., Abdullahi, S. M., and Khan, A. N. (2022). Deepfake detection for human face images and videos: A survey. *IEEE Access*, 10:18757–18775.

Minsky, L. and Aron, D. (2021). Are you doing the swot analysis backwards? *Harvard Business Review*.

Mirsky, Y. and Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Comput. Surv.*, 54(1).

Molvig, O. (2023). Talk to einstein.

Mori, M., MacDorman, K. F., and Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics Automation Magazine*, 19(2):98–100.

Office, U. C. (2022). *Copyright Law of the United States and Related Laws Contained in Tıtle 17 of the United States Code*. The United States Copyright Office.

O'Sullivan, D. (2023). Nonconsensual deepfake porn puts ai in spotlight.

P, S. and Sk, S. (2021). Deepfake creation and detection:a survey. In *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 584–588.

Prabhat, Nishant, and Kumar Vishwakarma, D. (2020). Comparative analysis of deep convolutional generative adversarial network and conditional generative adversarial network using hand written digits. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1072–1075.

Project, D. M. L. (2022a). *California Right of Publicity Law*. Berkman Klein Center for Internet and Society at Harvard University.

Project, D. M. L. (2022b). *New York Right of Publicity Law*. Berkman Klein Center for Internet and Society at Harvard University.

Project, D. M. L. (2022c). *State Law: Right of Publicity*. Berkman Klein Center for Internet and Society at Harvard University.

Project, D. M. L. (2022d). *Using the Name or Likeness of Another*. Berkman Klein Center for Internet and Society at Harvard University.

Rafezi, Z., Eskandari, H., and Saeidan, A. (2019). Designing a serious game "events" and investigating the effectiveness in modifying confirmation bias: A single subject study. In *2019 International Serious Games Symposium (ISGS)*, pages 89–93.

Rana, M. S., Nobi, M. N., Murali, B., and Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE Access*, 10:25494–25513.

Shi, H., Wang, X., Li, G., and Zhang, H. (2011). Fmea-based control mechanism for embedded control software. In *2011 International Conference of Information Technology, Computer Engineering and Management Sciences*, volume 1, pages 110–112.

Sinnott-Armstrong, W. (2022). Consequentialism. In Zalta, E. N. and Nodelman, U., editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2022 edition.

Stanojević, D. and Ćirović, V. (2020). Contribution to development of risk analysis methods by application of artificial intelligence techniques. *Quality and Reliability Engineering International*, 36(7):2268–2284.

Strubell, E., Ganesh, A., and McCallum, A. (2020). Energy and policy considerations for modern deep learning research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13693–13696.

Suwajanakorn, S., Seitz, S. M., and Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: Learning lip sync from audio. *ACM Trans. Graph.*, 36(4).

Suzuki, M. and Yamamoto, Y. (2021). Analysis of relationship between confirmation bias and web search behavior. In *Proceedings of the 22nd International Conference on Information Integration and Web-Based Applications amp; Services*, iiWAS '20, page 184–191, New York, NY, USA. Association for Computing Machinery.

Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395.

Thompson, N. C., Udrescu, S.-M., and Zhang, S. (2022). Here's how to make deep learning more sustainable. *IEEE Spectrum*.

Vamsi, V. V. V. N. S., Shet, S. S., Reddy, S. S. M., Rose, S. S., Shetty, S. R., Sathvika, S., M. S., S., and Shankar, S. P. (2022). Deepfake detection in digital media forensics. *Global Transitions Proceedings*, 3(1):74–79. International Conference on Intelligent Engineering Approach(ICIEA-2022).

Vincent, J. (2021). Tom cruise deepfake creator says public shouldn't be worried about 'one-click fakes'. *The Verge*.

Wickramasinghe, C. S., Marino, D. L., Grandio, J., and Manic, M. (2020). Trustworthy ai development guidelines for human system interaction. In *2020 13th International Conference on Human System Interaction (HSI)*, pages 130–136.

Wynn, N., Johnsen, K., and Gonzalez, N. (2021). Deepfake portraits in augmented reality for museum exhibits. In *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 513–514.

Zeitchik, S. (2022). Ready or not, mass video deepfakes are coming. *The Washington Post*.