# Preparing for the World of a "Perfect" Deepfake

Catherine Zeng, Rafael Olivera-Cintrón

Fall 2019

## Contents

# 1 Executive Summary

According to research out of cybersecurity company, *Deeptrace*, the number of "deepfake" videos on the internet have doubled in just nine months from 7,964 in December 2018 to 14,698. Of these "deepfakes", 96% were pornographic, often with a the face of a celebrity morphed onto the body of an adult actor engaged in sexual activity [1]. Accordingly, Facebook has invested $10M into a research effort to produce a dataset and benchmark for detecting deepfakes, and is partnering with top research institutions such as MIT, UC Berkeley, and Cornell Tech [2]. It's clear that deepfakes are alarming and firms like Facebook are doing something about it. But what are they? And why are they alarming?

The deepfake is the newest of tools coming from a long history of doctored media. Due to the increased concentration of users around social media and democratization of the means by which deepfakes are produced, the web is seeing an increasing propagation of hyper-realistic deepfaked videos; applications like the free and accessible *FakeApp* that enable you to make deepfakes without technical understanding of machine learning, and their increased realism and scale is largely due to improvements in the organization of datasets being fed into machine learning algorithms, as well as the introduction of Generative Adversarial Networks (GANs).

The GAN architecture is particularly suited for generating deepfakes because the architecture is optimized for producing fake content that fools itself; the system contains two parts: the generative part generates fake data that looks like the training data, and the discriminative part evaluates the generated data to see if it is fake or not; the system is rewarded when the generative part successfully generates content that fools the discriminative part. When trained on larger and more organized datasets, GANs can generate hyper-realistic deepfakes very quickly.

When truths are indistinguishable from falsehoods, we put at risk our democracy, our national security, and our public safety. When in the world of the "perfect" deepfake, the waters of fact and fiction are muddled, creating a fog of questioning what's real and what's fake, even when obvious. Politicians may use deepfakes to deceive the public for cheap political points. In which case, can our election process be trusted? In Gabon, it almost led to the upheaval of the entire government [3]. How might deepfakes make us question our national security in times of war? Deepfakes sent from adversaries can show our soldiers killing civilians to invoke an environment of distrust and instability. Deepfakes will complicate all rules of engagement, and will also create sticky situations domestically. Without proper accessible tools to identify deepfakes, people will not feel safe because if the deepfake were to victimize them, there would be no path towards

vindication. Because of how threatening and powerful those capacities are, we are proposing the government intervene and host a free and open channel where the public can access deepfake identification technology in order to curb the effects of disinformation.

Many prominent legal scholars, such as Danielle Citron and Robert Hansen, share the general consensus that existing laws are not equipped yet to handle a foreseeable future of deepfake proliferation, and until there is an attempt by the legislature to rectify that, analogous judicial precedent can be used to recommend how the Courts could interpret deepfakes. For example, defamation is historically difficult prove in court because it puts the burden of proof on the plaintiff. The plaintiff must be able to prove "reckless disregard" (of the truth) or "intentional malice." However, in the realm of deepfakes, would it be difficult to prove "intentional malice" when creating a deepfake that harms someone requires both intent and likely entails some sort of foresight of potential harm done? The Court doesn't normally like to establish strict guidelines on speech, but will generally support the plaintiff if "intentional malice" can be proved in cases involving defamation, intentional infliction of emotional distress, misappropriation of name or likeness, and false light. Therefore, if deepfakes used to harm can arguably be characterized as both "intentional" and "malicious," Courts would favor protecting people victimized by deepfakes. The Courts, however, often review free speech suits on a case by case basis considering all details of the trial. Therefore, it is uncertain whether case law will be primarily protect protect victims of deepfakes or favor the free speech rights of deepfake creators.

Although deepfakes are creepy, there are a variety of accurate tools that work well to detect them. For example, techniques for identifying deepfakes based on visual artifacts can already achieve accuracy rates of up to 86.6% [4] on videos from the FaceForensics dataset [5]. We propose that in order to address deepfakes, the government host a deepface detection platform that benefits individuals through an easy-to-use web application and corporations through a free REST API. This deepface detection platform should be built using a pipeline of deepfake detection techniques that together make it very difficult for a deepfake to slip through undetected.

We further propose that the government open-source the code base used for developing the deep fake detection platform with the primary arguments that an open-sourced code base is more cost-effective, enables faster iterations, and allows for algorithmic transparency. To make these arguments, we are strongly inspired by the open-source environment that enabled the development and maintenance of the Linux kernel, as documented in the text *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary* [6]. We believe that by open-sourcing the code base, the government-hosted deep fake detection platform would be more sustainably maintained in the future.

# 2 What are Deepfakes?

Deepfakes, named by combining terms "deep-learning" and "fakes", are machine learning generated audio or visual media that are loosely indistinguishable from real media; often, they replace a person's face in an image or video with someone else's face or other likeness. Due to the nature of this kind of content, deepfakes have been used to generate pornographic videos of celebrities, fake news and hoaxes, and even financial fraud in accounting [7]. In Gabon, the use of deepfakes even led to the attempted coup of their government[3]. As a result, deepfakes are widely considered a large threat to democracy, national security, and public safety.

Deepfakes are made using autoencoders and more recently generative adversarial networks (GANs), a category of machine learning models that have dramatically improved the realism of generated media. A recent paper released by Nvidia, the dominant producer of graphics processing units (GPUs) used for training and testing machine learning models, titled *A Style-Based Generator Architecture for Generative Adversarial Networks* [8] gained widespread press attention when images of people generated by its GANs were eerily realistic. Nvidia's model was trained on the CelebA-HQ dataset, a richly annotated dataset containing more than 200,000 celebrity faces in large pose variations and background clutter; and a new dataset they created called FlickrFaces-HQ (FFHQ) that includes 70,000 1024x1024 resolution images with even more variety than the CelebA-HD dataset in terms of accessories and backgrounds. As a result of the dataset's diversity, Nvidia's GAN model outputs contain flexible fakes varying in eyeglasses, hats, hair type, face shape, expressions, and backgrounds.

Deepfakes are have since become ubiquitous on the internet such that "deepfake artists" have emerged. For example, the user Sham00k runs a YouTube channel containing videos that superimpose Tom Selleck as Indiana Jones and Will Smith as Neo from *The Matrix* [9]. One of Sham00k's most popular videos was re-posted on actor and impressionist Jim Meskimen's channel, superimposing the faces of people Jim was imitating to his own in striking detail [10]. Comments left on the video show that the reaction to the effectiveness of deepfakes is split between awe and fear: "51% Impressive, 49% terrifying." - TheBoredEngineer; "1. This is really cool, 2. Society is totally screwed." - Zazz Razzamatazz.

## 2.1 History of Doctored Media

Doctored images and videos are not new and has been around ever since the conception of media. Matthew Brady, one of the earliest American photographers known for his depictions of

the Civil War, was also a successful photo manipulator. In the following group portrait of William Tecumseh Sherman and his top officers, for example, Matthew Brady edited Francis P. Blair into the photograph, a critical figure for defeating Confederates, to the far right [11].

Figure 1: Left: original photograph, Right: Francis P. Blair is added.

Often, images are even altered in order to create a better presentation for the audience. For example, the Pulitzer Prize-winning photograph of Mary Ann Vecchio kneeling over the body of Jeffrey Miller, who was killed by National Guardsmen during a protest against the war in Vietnam, was modified to remove a fence post behind Vecchio when the image was published in Life magazine [11].

Figure 2: Left: original photograph, Right: Fence post is removed.

Although doctored media is not new and has been documented in society for at least a century, two recent changes have made society significantly more vulnerable to deepfake manipulation: the increased concentration of users around social platforms on the internet, and technological improvements involving machine learning that have enabled deepfakes to be produced with realism at scale.

## 2.2 Danger of Deepfakes

### 2.2.1 Media Propagation and Large-Scale Distribution

According to Pew Research Center's 2016 update, 79% of Americans use Facebook, 32% of online adults use Instagram, and 24% of internet users use Twitter [12]. The extensive reach of the internet and the concentration of users around social media platforms has made it easier to spread false content to gullible ears.

For example, in November 2018, White House press secretary Sarah Sanders shared a video on her personal Twitter of Jim Acosta, a CNN reporter, in which he appeared to have treated a female White House intern aggressively by chopping her arm when she reached for his microphone [13]. This video was used to score political points against CNN. However, it was revealed by manual analysis that the video Sanders shared originated from Paul Joseph Watson, known for his conspiracy theory videos on the far-right website Infowars. This video was further revealed to be doctored so that Acosta's tussle for the microphone was sped up to give the appearance that he was chopping Sander's arm, and taken out of context as Acosta's statement "Pardon me, ma'am" was not included. Sander's video gave the illusion Acosta was particularly aggressive and shaped the political atmosphere unfairly against Acosta.

In the example above of Sarah Sander's tweet of the Jim Acosta doctored video, Sarah received over 66.5K likes, 18.6K retweets, and 52.7K replies. Further, when White House Press Secretary Stephanie Grisham, representing Trump's administration, shared the doctored video, she received 94K likes, 28K retweets, and 83K replies [14]. Therefore, the doctored video was able to reach over 4.7M people given a back-of-the-envelope estimation that assumes each retweet reach 100 additional people (in reality this number is most likely higher). This above example demonstrates how the increased concentration of users around social media platforms elevates the danger of deepfakes because it enables the mass distribution of content on platforms that users assume to be trustworthy.

### 2.2.2 Technology Enabled Scale and Realism

Technological breakthroughs in machine learning have enabled deepfakes to be produced at increasing realism and scale. Specifically, these breakthroughs were made possible by two advances: the emergence of large datasets, and development of Generative Adversarial Networks (GANs), a category of machine learning models developed by Ian Goodfellow et al. [15] where neural networks compete against each other in an unsupervised way; Yann LeCun, a Turing

award winner and the developer of convolutional neural networks, notably described GANs as "the coolest idea in machine learning in the last twenty years" [16].

As explained in Section 2.1, doctored images and videos are not new. However, the danger of fake media has grown because of the GAN's ability to generate doctored, and very realistic, media at scale using large datasets. For example, the graphics editing software Adobe Photoshop has existed since 1988, but an image of a fake face akin to one generated by Nvidia's GAN architecture trained on the CelebA dataset would take a professional photoshop artist multiple hours to make, not to mention downright impossible due to the level of detail required. In contrast, a trained GAN can make hundreds of photos in a second, with varying detail including expressions and angles for the same generated face.

## 2.3 Technological Advancements

### 2.3.1 Large Datasets

In August 2019, Scale API, a company led by 22-year old MIT dropout Alexandr Wang raised $100 Million dollars to become Silicon Valley's latest unicorn worth over a billion dollars [17]. How does Scale API make money? Companies provide Scale with data through their Application programming interface (API) and Scale labels the text, audio, pictures, and videos so that their client's machine learning models can be trained using the labeled dataset. Scale API's enormous success and profitability highlights the heavy dependence of machine learning models on the quality of the datasets with which they are trained on.

The development of increasingly realistic machine learning models has risen in tandem with the development of increasingly organized datasets, such that Wikipedia even has a page called *List of datasets for machine-learning research* [18] with datasets in categories such as image, sound, and text data and subcategories such as (within the image category) facial recognition, action recognition, and handwriting. In the case of Nvidia's *A Style-Based Generator Architecture for Generative Adversarial Networks* [8], in order to generate better deepfakes, Nvidia even created their own dataset FlickrFaces-HQ (FFHQ) so that they can have greater variation for their training data than the CelebA-HQ dataset in terms of age, ethnicity, image background, and accessories such as eyeglasses, sunglasses, hats.

### 2.3.2 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) are a category of machine learning models developed by Ian Goodfellow et al. in 2014 [15] where two neural networks compete with each other in an unsupervised way. GANs consistent of two parts: the generative network, which generates candidates, and the discriminative network, which evaluates them. The goal of the generative network is to generate datapoints that are similar to datapoints in the training set, and the discriminative network's goal is to judge these datapoints. The GAN's overall goal is to increase the error rate of the discriminative network: to "fool" it with generated datapoints that seem real.

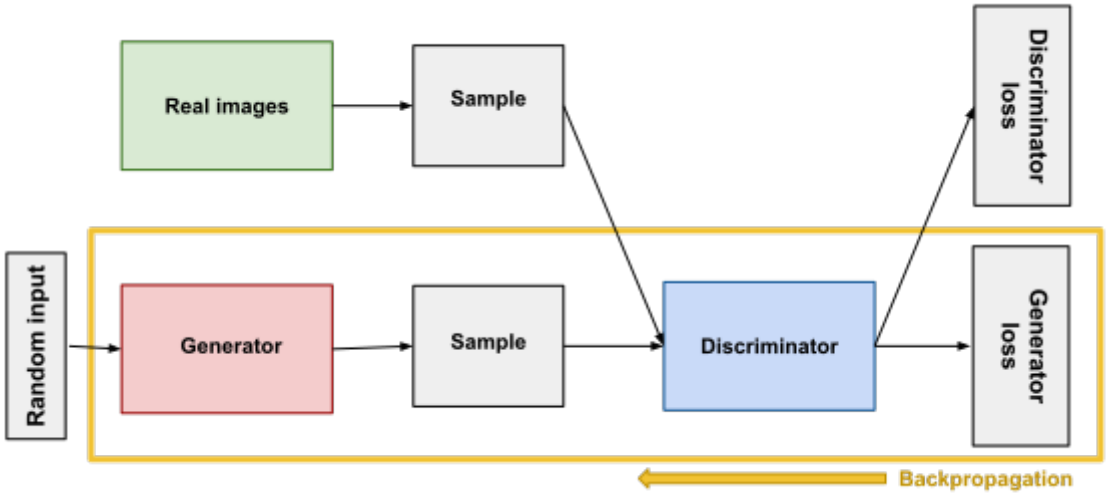


Figure 3: Block diagram of Generative Adversarial Network [19]

Given the architecture of GANs, it is easy to see how GANs are particularly suited for producing deepfakes; the model is literally optimized for producing new data with similar features to the model's training set. In the case of deepfake images of faces, a GAN model trained on the highly rich Celeb-A dataset is optimized to generate pictures as close to the original dataset as possible. It's easy to see how abuse of GANs technology can produce frightening consequences that blur the line between real and fake images.

Nonetheless, it is worth pointing out that GANs are not together evil. They are useful in a variety of legitimate fields including:

- **Video game modding** - In order to up-scale low resolution images from old video games that retain original levels of details while producing sharper textures, GANs have been employed in games such as Final Fantasy VII and The Legend of Zelda: Twilight Princess [20].
- **Astronomy** - In order to improve astronomical images where random and systematic noise

from the sky background is common, GANs are employed to recover features from artificially degraded images and detailed features [21].

### 2.4 Introduction to our Recommendation

Fortunately, we are not yet in the world of a "perfect" deepfake. Deepfakes can be detected with an accuracy of up to 86.6% using some methods (refer to section 5). While technology like that exists, we believe that where the government normally stands aside and lets media and people handle the propagation of disinformation, for the case of deepfakes, the government needs to have an active role in making accessible the technology to identify deepfakes. We see deepfakes as a threat unlike any seen in the realm of disinformation. Applications like FaceApp make deepfakes easily accessible, democratized, and possible without heavy computing power. If there is no reliable way for the public to check deepfakes, the possibility of them will always play a subliminal, yet significant role in directing the the public discourse. What is real and what is not when anything real could be portrayed as fake and anything fake could be portrayed as real? Therefore, we suggest the government make these tools accessible to people by hosting them on a web application, and accessible to media platforms through an API (section 6). The public will only be safe when the tools to identify deepfakes are just as accessible as the tools to produce deepfakes (assuming they're on par with each other).

## 3 Threats to Democracy, National Security, and Public Safety

The exponential progress of technology will one day bring about the perfect deepfake – an AI generated video so real that it is completely indistinguishable from reality. Would a perfect deepfake detector accurately identify a perfect deepfake? Does the immovable object stop the unstoppable force? The lack of answer to that question creates a cornocopia of concerns, because although it may not seem like it, the world of the *perfect deepfake* is one which is inevitable and one which we must be prepared for.

If we cast aside these concerns and leave them for another day, we will soon be faced with immeasurable grievances. Because once a lie is indistinguishable from truth, once facts are indistinguishable from fiction, we will not only lose trust in each other, but transitively we will begin to see the corrosion of our democratic process, the emergence of serious formidable national security challenges, and infinitely occurring threats to the safety and well-being of the public.

## 3.1 Democracy

The sanctity of democracy pivots around the idea of open and verifiable information. Can this idea still be embraced if we remain unprepared in the world of a perfect deepfake? In just 2016-2019, the United States has witnessed some of the extents of *old doctored media*. In May 2019, video of a press conference of Speaker of the House, Nancy Pelosi, was doctored by Shawn Brooks in a way that made her appear to be slurring her words in a drunken way [22]. Brooks achieved this by "slowing Pelosi down without lowering the pitch of her voice" [22]. Out of context, to the average observer, this is just a harmless joke, but it made the news because President Donald Trump tweeted it out as if it was true and his press team began speaking in interviews as if it was reality [23] [22]. This video was quickly revealed to be fake, but the damage had been done in the eyes of millions who had already seen the post. Most recently, in December 2019, the Joe Biden 2020 Presidential campaign released a television advertisement accusing President Trump of being a laughing stock on the world stage. The only problem is that the ad was deceptively edited to create the illusion that during a speech President Trump gave to the United Nations, the audience laughed immediately after a claim that his "administration had accomplished more than almost any administration in the history of our country" [24]. In the actual speech, there was a delay between the delivery of that line and the laugh, not to mention the laugh was likely prompted by something else President Trump said soon after that line. If even just these manually doctored videos are enough to be effective political fuel against opponents, what foulplay is to be expected as we approach the perfect deepfake?

The day is October 29, 2004, days before the Presidential general election, Al-Jazeera receives a video tape to their office in Pakistan of terrorist leader, Osama bin Laden, claiming responsibility for the terror attacks of September 11, 2001. The nation is galvanized. Presidential incumbent, George W. Bush, capitalized on public concerns of national security and successfully painted himself as the only candidate adequately prepared to fight radical Islamic terrorism, giving him the edge and winning the election days later [25] [26]. Now, let's assume the year is 2028, completely hypothetical Republican Presidential candidate, *Daniel Tucker*, is running against completely hypothetical Democratic Presidential hopeful, *Lucas Ron*. Daniel Tucker believes that the election is too close for his liking so he, the 'pro-military' conservative, having noted the events of 2004, devises a plan using Republican sponsored, military-grade deepfake technology. Days before the 2028 election, a video of an unnamed cyberterrorist leader claiming responsibility for the nation-wide, devastating power grid failure of 2025 and promising more attacks, is leaked to the public. Daniel Tucker uses this moment to take a firm stance on cybersecurity and the protection of American interests in the 'realm of cyber.' The public, panicked by this immediate threat,

tips their support over to Daniel Tucker. Daniel Tucker, having orchestrated the entire predicament, edges out his opponent to win the 2028 presidency under public consensus that he will be stronger against this newly revealed enemy. Daniel Tucker now has a puppet opponent – the fake cyberterrorist group – to continue scoring victories off of whenever he wants cheap political points. This kind of hypothetical election manipulation situation may seem silly, but it is far from being out the realm of possibility.

The mere fact an event like this is in the scope of what could happen means we need to take immediate steps as to not undermine our democracy. So far, California has been the only state to ban the use of deepfakes in political campaign promotions within 60 days of an election, enough time for fact-checkers to label a deepfake and for a candidate to set straight any confusion [1]. How would that even be enforced though? What resource should be made available to distinguish between real and deepfaked videos? What is to stop these "eve-of-election" election smears?

Now, imagine instead a scenario created by his opponent Lucas Ron, leaking a deepfake video of Daniel Tucker partaking in some sort of criminal activity which compromises his character. What if Daniel Tucker actually partook in criminal activity but claimed it was a deepfake created by Lucas Ron, when it was not?

What can be trusted about our political process if any information can be produced by a deepfake? Political candidates will be able to get away with the most bizarre things by claiming that the capturing of that activity was in it of itself a deepfake – a phenomena first referenced by Danielle Citron and Robert Chesney as the "Liar's Dividend" [27]. Donald Trump was savvy to this in 2016 insinuating, soon after a tape of him saying incredibly obscene things was leaked, that the tape was not in fact real and it was doctored [28]. This claim never stuck, but in the world of a perfect deepfake, it very well could. Even more terrifying is the deepfake which nearly brought the end to the current Gabonese government. In Autumn of 2018, Gabonese President Ali Bongo was receiving medical treatment in Saudi Arabia and London. He kept this a secret from the public because had he been deemed unfit to continue being president, he and his family would lose their 43-year long claim to the presidency under their constitution. The public was very confused by his lack of public appearance, especially amidst the rumors [3]. Could the rumors be real? Could the president be sick or worse? Fortunately, the President's advisors revealed he would give his customary New Year's address. The only problem being that the glassy eyes and unusual face perturbations indicated the address was likely the product of a deepfake [29]. It seemingly was enough to trick the public, but the military, being suspicious of this video, attempted an unsuccessful coup [3]. In only the early stages of deepfake technology, a coup has

[1] California Assembly Bill No. 730

already been attempted. The problems deepfakes create for our democracy go far beyond keeping politicians accountable and ensuring fair elections. Misuse of deepfakes could quite literally topple governments.

## 3.2 National Security

The concept of the "Liar's Dividend" extends not only to politicians but to everyone. In 2010, with the help of whistleblower Chelsea Manning, Wikileaks uncovered footage of United States air crew launching an air strike on two waves of civilians and then laughing about it [30]. Some of those men were held accountable for their actions, but had we lived in a world of deepfakes, those men could have made the claim that the videos themselves were faked by Chelsea Manning, who was (and is) already being treated as a traitor for having leaked the footage. In a similar, yet completely opposite line of thinking, a major problem could arise with innocent soldiers being deepfaked onto videos of the military killing civilians. Not only would that needlessly create further distrust between the public and the military, it also illustrates the dichotomy created by the Liar's Dividend. The dichotomy between those who lie about being victimized versus the people who are actually victimized exacerbates the already difficult process of distinguishing truth from fiction in a world of perfect deepfakes.

Naturally, the more complex national security created by deepfakes are the more threatening. Deepfakes introduce many new avenues by which nations can partake in geopolitics, and turns diplomacy into a much more interesting game. The date is January 1917, the British, seeking the help of the United States, conveniently intercepted the Zimmerman Telegram, a message sent to the German Minister to Mexico, offering United States territory in return for alliance with the Germans in World War 1 [31]. The Mexican government declined, but this event signified the precursor to the United States entering the First World War. In the world of the perfect deepfake, the year is 2026, the Indians are standing on the sidelines of a war brewing between the United States and China in what is being call the Third World War. Suddenly, the United States government, having only their alliance with Japan and Australia to supplement their lack of presence in that region, intercepts a message which they quickly deliver to the President of India from the President of China intended for the President of Pakistan. The message is one seeking some sort of alliance, playing on pre-exsisting Pakistan-India conflict to request an invasion of India with the intent to expand the battleground further west with the promise on returning that land to Pakistan after the war. India, galvanized by this threat, having initially chosen to remain neutral in World War 3, decide to partake in the conflict and preemptively invade Pakistan and post troops outside the border of China while joining the alliance with the United States. The surprising es-

calation by India caught the world off-guard and left China in a precarious position – cornered from their Western, Southern, and Eastern front. The tides of the war shifted. What India was unaware of is that the United States had actually deepfaked the video of the President of China with the intention of deceiving the Indian government into being a strategic ally in the region.

During the Presidency of Donald Trump, the idea of foreign interference in United States elections has raised important concerns involving foreign influence in choosing representatives. Although the Russians did not directly meddle with our electronic ballot system, the evidence does show that they largely enabled "troll farms" to post subliminally manipulative content that would influence public perception of who they wanted elected [32] [33]. If instead of memes, the troll farms were actively producing deepfakes of our political candidates saying or doing things which are not representative of their intentions, we would find ourselves in significant trouble. Whereas media and people were once the ones with carrying onus to expose fake news and doctored media, the sophistication of the deepfake requires for the government to intervene and provide the tools necessary to identify deepfakes accessible and free. Deepfakes are increasingly easy to create and at the same time they're increasingly becoming more realistic. Because of this, a potential foreign troll farm, planting deepfake media of all kinds would muddle the water between real statements made by candidates and fake candidates, and could thereby easily control the narrative in favor of their preferred candidate. This fog of disinformation could be pacified with the introduction of a transparent, free resource, hosted by the government and community driven, with the purpose of using the techniques elaborated on in section 5 to distinguish between real and deepfaked content uploaded by members of the public and also accessible through an API for corporations seeking to auto-label video uploads on their media platforms. Granted, the current methods of identifying deepfakes are not full-proof yet, but even a limited step forward is a step forward when it comes to providing a sense of security to people about the validity of their information.

### 3.3 Public Safety

Without accessible tools to identify deepfakes, all people are at risk to be the victims of a deepfake with little possibility to vindicate themselves. Beware, deepfakes will create public discomfort and distrust of all information, not just that which is coming from the top, but also that coming from each other. When you can't trust your fellow brothers and sisters what kind of damages are done to society?

Public figures are the most likely ones to suffer. Stock prices of companies going down because deepfaked videos of executives surfaced of them drinking while driving. Business nego-

tiations failing because executives of one company receive deepfaked video of the executives of the other company speaking poorly about them and in a superior tone. Celebrities that don't consent to their likeness being used for advertisements or modeling. Celebrities deepfaked onto pornographic material.

Pornographic content using the faces of celebrities like Natalie Portman, Emma Watson, and Gal Gadot are becoming increasingly more common [34]. Although public figures are mainly targeted by this type of deepfake, the general public should not feel too detached from the problem. Don't feel apathetic. Deepfakes can be made by anyone with very accessible computation power. What is currently primarily affecting celebrities will come to haunt the general public in new avenues of revenge porn, cyber bullying, and blackmail.

Exiting relationships is already a challenge for people, but if we are unprepared to provide accessible means to identify deepfakes, vengeful partners who might not even have pornographic material of their ex-partner and just a lot of pictures can create any kind of situation they want. This could be detrimental to someones social life, their work life, and their family life. For example, imagine a hypothetical Jacklyn Johnson who broke up with hypothetical boyfriend of 4 years Rob Porter. A week later Jacklyn's mom is sent a video of Jacklyn having sex with a random stranger and told that this was the reason which the break up occurred. All of Jacklyn's friends and all of Jacklyn's co-workers also received the same video. What happens in a world where Jacklyn can't immediately put the video into a filter and show everyone that it is fake? Do revenge porn laws even consider deepfakes? The California State Senate recently passed AB-602, which protects people's faces from being put on pornographic content with use of deepfakes [2], but not everyone lives in California. Not to mention, regardless of Jacklyn's ability to press charges, her reputation maybe irreparably damaged with people in her family and friends who have no reason to believe the video is falsified.

Remember how challenging high school was? If bullying revealed to people how cruel children could be to themselves, imagine what that could look like in a world where you can make a video of anybody doing anything. In a high school where being gay is still stigmatized, someone could leak a video of one of the theatre kids being romantic with another boy and bully him for it, even if he might not even be gay. Popular girls, stereotypicaly ruthless among themselves, releasing deepfaked nude pictures of someone they're trying to cancel. Administration who find themselves disliking a particular student and then report a video of him partaking in drugs in the back of the school, where that person could otherwise pass a drug test. The new avenues of bullying and cyberbullying that deepfakes create is terrifying and despicable.

[2]California Assembly Bill AB-602

Another nightmare of deepfakes is a scenario where a hypothetical average guy, Tim Putter, receives electronic mail with tapes of him doing particularly obscene things and told that in order that to go away he'd need to pay a ransom, or commit a crime, or do another obscene thing. Tim Putter knows that is not him, but without a way to validate the status of deepfake of the content, Tim might go to lengths to make sure others do not believe he did those things. Blackmail and extortion, when the hidden villain can realistically not be tracked by local law enforcement. What is to be done then? Imagine a situation where none of this content is even obscene. The villain of the story just creates a video of Tim Putter entering an ambulance on a stretcher and sends this to an unsuspecting mother or grandparent.

# 4 Analogous Judicial Precedent

The emergence of the "deepfake" creates a new environment of content, one that along with creating new tools to benefit varying forms of expression, also brings a cornucopia of evils. Robert Chesney and Danielle Citron in "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security" outline this dichotomy. Deepfakes will one day enable a future of new forms of educational experiences, as well as artistic expression in the form of satire or re-enactments, and might even allow a new space for people with disabilities to have experiences they otherwise could not [27]. However, this same advancement also provides new tools for exploitation in the form of blackmail, sabotage in the form of impersonation, and harm to society measured in distrust of news, discord among political and social groups, and further distrust of our political discourse, diplomacy, and democracy.

There are many different avenues and lanes of legal recommendation coming from the legal scholars and deepfake experts currently advocating for these issues. Whereas Danielle Citron and Robert Chesney have suggested narrowing the language of Section 230 of the Communications Decency Act (CDA)[3] to create an environment where the "Good Samaritan" provisions would incentivize media platforms to remove deepfake content [27][35], Richard Hansen of University of California, Irvine, instead believes that there exists Constitutional framework to mandate "truth-in-labeling law" which would require media platforms to use the best available technology to label doctored content but not remove it [35]. Within that premise, we agree completely with Hansen, and disagree with Citron and Chesney who seem to suggest that moderation of content must mean complete removal of content. Our ultimate proposal is providing a government hosted service by which media platforms can access an API that auto-labels uploaded content as

[3]Communications Decency Act of 1996

real or fake using the identification techniques described later in section 5. Hansen also believes there might be some avenue to regulate deepfake contents through the existing election laws, but notes that there are a variety of precedents established in the Supreme Court about violations of the First Amendment in reference to laws penalizing, what Hansen refers to as, "false election speech" [35].

Cases such United States v. Alvarez [4] exemplify how diverse the opinions within the Supreme Court can be when regarding deceptive information produced by candidates of political elections. Though, it is possible that through cases like Minnesota Voters Alliance v Mansky [5], verifiable falsehoods – such as where the location of a polling location is – can be justifiably removed from the public discourse, as patent falsehoods if removed do not equate political censorship. However, as Citron and Chesney note, the Supreme Court in Brown v Hartlage [6][27] stated that the "State's fear that voters might make an ill-advised choice does not provide the State with a compelling justification for limiting speech." Helen Norton, a free speech scholar, suggests in Lies and the Constitution [36], that regulating false election speech might further agitate fears of "government overreaching and partisan abuse." Douglas Harris, fearing the "False Pornography," believes that through a "federal criminal statute" (assuming the content creators can be identified) published deepfakes should be banned whereas creators should be protected in creating personal deepfakes[37]. Although, the scholars differ in their interpretations of the problems and paths towards a solution, there seems to be consensus among scholars like Harris, Citron, Chesney, and Hansen that existing laws are not currently equipped to adequately protect the victims of deepfakes.

The sophisticated challenges arising from this new form of content will be faced with equally challenging policy questions and must be met with equally sophisticated policy solutions. If the answer is not to ban deepfakes, like the Cyberspace Administration of China (CAC) practically did [38], because it is infeasible to track all individuals producing deepfakes because of the democratized and accessible nature of the tools to create deepfakes. If the answer is to allow the good of deepfakes, but possibly regulate the bad, then what existing legal precedents can be used to protect individuals assuming the creators of the deepfakes can be identified? There clearly exist some shared trains of thought as well as diverging ideas when it comes to what should be done about this issue, but if the legislature passes no new laws related to creating deepfake protections, scholars and the public will need to wait until a court case establishes some kind of judicial precedent. Fortunately, there exists judicial precedent in many analogous cases.

[4] United States v. Alvarez, 567 U.S. 709 (2012)
[5] Minnesota Voters Alliance v. Mansky, 585 U.S. (2018)
[6] Brown v. Hartlage, 456 U.S. 45 (1982)

## 4.1 Defamation

The simplest definition of defamation is an expression which injures the reputation of another. Within the United States court system, the plaintiff of a defamation case must initially prove [39]:

1. A false statement purported to be fact
2. Publication or communication of that statement to a third person
3. Fault amounting to at least negligence
4. Damages, or some harm caused to the person or entity who is the subject of the statement.

Defamation in written form is "libel," while in spoken form it is "slander". The New York Times Co v Sullivan [7] established the standard of "actual malice," where the plaintiff would need to "demonstrate the publisher's knowledge that the information was false or that the information was published with reckless disregard of whether it was false or not." *New York Times Co v Sullivan* determined that advertisements, regardless of containing false content, could only be classified as defamation if there is presentable evidence of "actual malice." Curtis Publishing Co v Butts [8] went on established the precedent that public officials could not sue for libel unless it could be proven that the information was published with malicious intent. Unlike other types of court suits, defamation does not weigh the burden of proof on the defendant, but instead mostly on the plaintiff. Due to this, defamation has been historically difficult to prove in court.

As a result of Hustler Magazine v Falwell [9], a case brought about by, public figure, televangelist, and political commentator, Jerry Falwell against the satirical Hustler Magazine, which portrayed him as an "incenstuous drunk," the Supreme court ruled that the First Amendment protects "statements that cannot 'reasonably [be] interpreted as stating actual facts' about an individual". In layman's terms, statements that are so ridiculous to be clearly not true are protected from libel claims. Given that, how might a case go where a plaintiff harmed by the creation of a deepafake has irrefutable proof of malicious intent, but the defendant claims the the deepfake was so clearly out of the realm of possibility that they deserve protection from a libel claim. In a this world of deepfakes, is there anything that is out of the realm of possibility? The court would need to expound what constitutes ridiculousness because of the climate created by deepfakes.

The creation of the Communications Decency Act of 1996 (CDA) [10] provided the defense necessary for American Online (commonly known as AOL) against Kenneth Zeran, who sued AOL

[7] New York Times Co. v. Sullivan, 376 U.S. 254 (1964)
[8] Curtis Publishing Co. v. Butts, 388 U.S. 130 (1967)
[9] Hustler Magazine v. Falwell, 485 U.S. 46, 108 S. Ct. 876 (1988)
[10] Communications Decency Act of 1996

for being the ISP by which a 3rd party created a hoax confabulating Zeran and glorification of the Oklahoma City Bombing while also providing his address. Whereas in previous cases like Stratton Oakmont v Prodigy [11] and Cubby Inc. v CompuServe Inc.[12], the Supreme Court ruled to penalize the host, the introduction of the CDA allowed for AOL claim not liable for the damage caused by others on their platform. Had Zeran been targeted by a deepfake attack how might've the case changed? In our estimation, deepfakes are such a substantial upgrade from typical disinformation that the government needs to provide ways by which platforms can label their content uploads automatically as real or deepfake. The government providing a standard of content moderation including provisions for deepfakes as painless to media platforms as implementing an API which automatically labels uploaded deepfake content, more formally outlines the space which media platforms can continue to avoid having liability from the content of its users. Media platforms benefit from accepting these terms because it removes any sort of ambiguity about what is allowed and what is not – what should be removed and what should not. Media platforms just need to inform the public of the content they're being shown. "You're looking at a textpost with two JPGs, one MOV, and four deepfakes!"

## 4.2 Intentional Infliction of Emotional Distress

Intentional Infliction of Emotional Distress (IIED) is characterized by the plaintiff intentionally creating severe emotional distress to the defendant through actions which could be defined as "extreme and outrageous" [40]. In order for a tort of IIED, a reputable presumption of [40]:

1. The defendant acts
2. The defendant's conduct is outrageous
3. The defendant acts for the purpose of causing the victim emotional distress so severe that it could be expected to adversely affect mental health
4. The defendant's conduct causes such distress

The precedent established in Snyder v Phelps[13] summarizes the general sentiments of the Supreme Court when it comes to cases of IIED. Matthew A. Snyder died in combat, and the Phelps family is the main family of the Westboro Baptist Church. The Westboro Baptist Church is a church located in Topeka, Kansas which practice a "fire and brimstone" fundamentalist religious faith [41]. The Westboro Baptist Church believes that because the United States tolerates

[11]Stratton Oakmont, Inc. v. Prodigy Services Co., 1995 WL 323710 (N.Y. Sup. Ct. 1995)
[12]Cubby, Inc. v. CompuServe Inc., 776 F. Supp. 135 (S.D.N.Y. 1991)
[13]Snyder v. Phelps, 562 U.S. 443 (2011)

homosexuality, God punishes the United States by killing soldiers. These beliefs and presuppositions lead the Westboro Baptist Church to protest outside the funeral sites of fallen soldiers a message opposed to American war and American acceptance of homosexuality. "God hates fags." Having experienced the hatred coming from this group, Matthew Snyder's father attempted to claim IIED. Unfortunately for him, the protesters were off-site and the majority of the message they were spewing was that of stances on issues and less of personal attacks. Because of the questionable, yet still mostly objective content being spewed by the protestors, the Supreme court ruled in favor of the Phelps's right to protest. The attacks were those of public concern, although their manner of protest was questionable, therefore they were protected by the First Amendment. However, the language of the case does seem to assert that the language and actions in question in an IIED tort must be reviewed on a case by case basis.

A case involving deepfakes and a plaintiff claiming IIED would likely fall under that same standard of review. For example, had the Phelps family rented a blimp with giant flat screen television and played a deepfaked video of Matthew Sydney killing civilians above the funeral with no mentions of general objections to war, the court may have ruled that the attacks were beyond just a matter of public concern. Whereas, had they played a deepfaked video of him killing civilians with messages condemning United States imperialism, the court likely would've ruled as they did in the original case.

### 4.3 Privacy Tort

"One who intentionally intrudes, physically or otherwise, upon the solitude or seclusion of another or his private affairs or concerns, is subject to liability to the other for invasion of his privacy, if the intrusion would be highly offensive to a reasonable person" [14]. In 1997, the Restatement (Second) of Torts Section 652 created the language necessary for the court to protect victims of privacy invasion. Within this document, the court made distinctions between four "privacy torts," by which any if violated, could constitute an invasion of someone's right to privacy:

1. Unreasonable intrusion upon the seclusion of another, as stated in 652B
2. Appropriation of the other's name or likeness, as stated in 652C
3. Unreasonable publicity given to the other's private life, as stated in 652D
4. Publicity that unreasonably places another in a false light before the public [15]

[14] Restatement of the Law, Second, Torts, Section 652
[15] Restatement of the Law, Second, Torts, Section 652

Using these principles and guidelines, privacy laws and more specifically data privacy laws, such as the notable California Privacy Rights Act (CCPA)[16], are able to be implemented to protect citizens and consumers. For the purpose of inspecting existing precedent involving cases that will likely be tangential or analogous to that of deepfakes, we have chosen to only focus on two and four – how the court has ruled in cases of appropriation of another's name or likes, and cases of false light, respectively.

#### 4.3.1 Appropriation of Name or Likeness

The unlawful and unauthorized use of someone else's name or likeness provides legal grounds to claim an invasion of privacy. Unlawful use of someone else's name or likeness is characterized by three elements [42]:

1. **Use of a Protected Attribute** - The plaintiff must show that the defendant used an aspect of his or her identity that is protected by the law. This ordinarily means a plaintiff's name or likeness, but the law protects certain other personal attributes as well.

2. **For an Exploitative Purpose** - The plaintiff must show that the defendant used his name, likeness, or other personal attributes for commercial or other exploitative purposes. Use of someone's name or likeness for news reporting and other expressive purposes is not exploitative, so long as there is a reasonable relationship between the use of the plaintiff's identity and a matter of legitimate public interest.

3. **No Consent** - The plaintiff must establish that he or she did not give permission for the offending use.

This tort exists to protect the value of someone's name or likeness, so it often may be more pertinent for famous individuals, and is also one which the types of protections available vary from state to state. Individuals, especially famous ones, have the right to be compensated for their likeness. In the case of Grant v Esquire, Inc., Hollywood star, Cary Grant, who had previously consented to appear in an Esquire article in 1946 about his clothing tastes, sued Esquire in 1971 because they repurposed his face from the 1946 pictures to put on a different model in 1971 with the purpose of drawing distinctions between old and new styles [17]. The Court did not find that there was actual malice or negligence of facts on the side of Esquire... The Court found that Esquire simply used his likeness without malice, and that because his likeness is likely worth some form of monetary value, Esquire would need to compensate Grant for that monetary value. The

[16]The California Consumer Privacy Act (CCPA)
[17]Grant v. Esquire, Inc., 367 F. Supp. 876 (S.D.N.Y. 1973)

Court deemed an appropriate financial settlement, but did not want to set any sort of precedent, because although the Court does not want to allow libelous publications, they definitely prefer to remain on the side of not “impeding untrammeled public debate”.

In an analogous hypothetical case, *Pamela v Fragrance*, Fragrance, a perfume company, reuses the likeness of Pamela from a prior photo-shoot to create a deepfake model used to promote their newest perfume line. *Grant v Esquire* and *Pamela v Fragrance* are parallel court cases. Because of the clear lack of "malice," the court would likely refuse to set any sort of precedent on speech and establish an appropriate financial settlement. However, *Grant v Esquire*, does seem to suggest that the likenesses of public figures are worth monetary value. So in a world of deepfakes, where celebrities are not consenting to their likeness being used on deepfaked videos, celebrities are likely entitled to financial compensation.

In the eyes of the Court, the concept of obscenity is too vague and heavy enforcement of it would inhibit free expression. Justice Brennan summarized this in Rosenbloom v Metromedia – “...the vital needs of freedom of the press and freedom of speech persuade us that allowing private citizens to obtain damage judgments on the basis of a jury determination that a publisher probably failed to use reasonable care would not provide adequate ‘breathing space’ for these great freedoms. Reasonable care is an ‘elusive standard’ that ‘would place on the press the intolerable burden of guessing how a jury might assess the reasonableness of steps taken by it . . .’ Fear of guessing wrong must inevitably cause self-censorship and thus create the danger that the legitimate utterance will be deterred" [18].

### 4.3.2 False Light

If an actor publishing information about another, knowing the information was false and acting with a “reckless disregard” as to the falsity of the publicized matter, and the victim had evidence of “actual malice”, the victim could then litigate the actor for placing him in a false light towards the public. False light is often times confused with defamation because both share the need to provide evidence of “actual malice” and the torts have similarly sounding expectations. Unlike defamation, false light compensates individuals for hurt feelings and not for hurt reputation. In order for a false light claim to pass it must pass four criteria:

1. The false impression would be highly offensive to a reasonable person
2. The actor knew the impression was false, or acted with reckless disregard as to the falsity of the publicized matter and the false light in which the victim would be placed” [43]

[18] Rosenbloom v. Metromedia, Inc., 403 U.S. 29 (1971)

3. The defendant publish the information widely

4. The publication identifies the plaintiff " [44]

Time, Inc. v Hill[19] establishes the need to provide proof of actual malice. The Hill family lived a traumatic and very bizarre experience, and Joseph Hayes wrote The Desperate Hours, which was eventually converted to a Broadway play. Life magazine, under the publisher Time, Inc., then wrote an article showing the comparisons between the on-stage performance and the accounting of actual events. The Hill family then sued claiming that their privacy was being invaded and their story was being made public simply to promote the Broadway performance. The Court ruled that Time, Inc. was within their rights invoking the lack of "actual malice," establishing precedent for that with all future false light cases (note: We estimate that the Hill family likely might've been able to seek financial settlement had they instead pursued the misappropriation of name or likeness privacy tort.). False light cases are generally approached on a case by case basis.

Alfred Hill wrote, "Thus the question is whether an invasion of privacy that is constitutionally protected loses that protection when accompanied by false statements uttered with the requisite degree of fault. To be sure, calculated falsehoods "enjoy no immunity" under the Constitution, as the Court said in Time, Inc. v. Hill; but it does not follow that there are no constitutional limits on the consequences which may be visited upon one who utters such falsehoods" [45]. If public access to the likeness of an individual available on infinitely many pictures on Google Images can be interpreted to be "an invasion of privacy that is constitutionally protected," then Alfred Hill might argue that when those images are fed into a GAN to produce a deepfake, a visual "false statement," then the initial "constitutionally protected invasion of privacy" loses its Constitutional protection. Therefore, "with the requisite degree of fault," deepfakes should always be considered "calculated falsehoods" based on the words of Hill. We agree and disagree. We believe the deepfakes that harm an individual were like proceeded with an intent to harm because of the deliberation during the creation of a deepfake. However, deepfakes can be used for artistic expression such as the the deepfake-like technology used in Forrest Gump to recreate and satirize historical events, as cited during the Congressional Hearing about deepfakes. Although both are technically "calculated falsehoods," we do believe these are important distinctions to make.

[19]Time, Inc. v. Hill, 385 U.S. 374 (1967)

# 5 Techniques for Identifying Deepfakes

Currently, techniques for identifying deepfakes can achieve incredible accuracy rates of up to 86.6% [4] on videos from the FaceForensics dataset, a benchmark of videos created for forgery detection [5]. We propose that by using a variety of these media forensics techniques through feeding them into a pipeline, we can build a tool that can identify most deepfakes on the internet.

## 5.1 Metadata Filter and Transparency

Image and video metadata is text information contained within a media file that includes production information such as date of creation, location of creation, camera used, and other details of relevance. Common metadata include: Information Interchange Model (IPTC), Extensible Metadata Platform (XMP), EXchangable Image File (Exif), Dublin Core Metadata Initiative (DCMI) and Picture Licensing Universal System (PLUS).

Although it is possible to alter media metadata, many images and videos on the internet often still contain the signature of editing programs. Therefore, it is easy to catch a number of doctored media through the initial metadata filter. A technique that would be useful for preliminary detection of deepfakes would be to reveal the metadata contained in the media file, as this can give clues to how the media was produced without expending computational resources.

## 5.2 Error Level Analyses (ELA)

For methods of lossy compression for digital images such as JPEGs, a common image type, Error Level Analysis (ELA) can be used to analyze compression artifacts [46]. For doctored images, the data may consist of areas containing different levels of compression artifacts due to being subjected to different levels of lossy compression. The compression ratio of particular portions of the image changes with respect to others; different variations in the level of compression artifacts means that an image has been doctored.

## 5.3 Visual Artifacts

Faces generated by deep fakes often have residual visual artifact that include one or more of the following: global inconsistency, inaccurate illumination estimation, and/or faulty geometry estimation [4]:

- **Global Inconsistency** - Although faces generated by deepfakes are supposed to support the interpolation of images, the mixture of different faces is not always consistent, so that they lack global consistency. For example, although heterochromia - the phenomenon of differently colored irises, is relatively rare for humans, photos generated from deepfakes often have high variances in color between the left and right eye.

- **Inaccurate Illumination Estimation** - Illumination, or the source of light and the shadows it generates on faces, is often distorted when transferring from the original image to the forgery. In *Face2Face: Real-time Face Capture and Reenactment of RGB Videos* [47], for example, the illumination and rendering estimations are modeled explicitly. However, in most deep-learning based models, illumination estimations are usually learned from the data implicitly. This often leads to imprecise estimations of incident illumination.

  An example of inaccurate illumination estimation is that in many deepfakes, shading artifacts can be spotted in one area of the nose where one side can be rendered unnaturally dark; it is hypothesized that limited illuminations models fail to take interreflections, which are observed on concavity or when multiple objects are located near each other, into account [4]. Further, reflection details are often missing in eyes generated through deepfakes. Specular reflections noticeable in real images are often unconvincingly generated in deepfakes; missing reflections are often simplified into a white blob, leading to a dull appearance of the eyes.

- **Faulty Geometry Estimation** - Similar to the case of illumination, geometry estimations are often made to fit a morphable model to images, and we can spot artifacts that arise from imprecise estimations of geometry, as seen by the data from the Face2Face model. Specifically, artifacts that show up as strong edges or high-contrast spots around the boundary of an overlaid face mask often appear on spots such as the nose, face, and eyebrows. Further, geometries such as teeth are often not modeled around; often teeth appear as a single white blob instead of as individual teeth.

## 5.4 Inconsistent Head Poses

Deepfakes are often created by replacing face regions of the original image with synthesized faces. This process often reveals errors when 3D head poses are estimated from head regions, as shown by the study *Exposing Deep Fakes Using Inconsistent Head Poses* [48].

During the process of swapping faces, landmark locations, which are locations on human faces corresponding to structures such as the tips of the eye and mouth, of fake faces are often different

from those in the original face. Because people have different facial structures, deepfakes do not guarantee that landmarks align. Often, deepfakes swap faces in the central face region, but the outer contours of the face remains the same while landmarks at the center are inconsistent from 3D head poses estimated from central and whole face features of the original image. Specifically, head differences between the central and whole face regions are small in real images but large in deepfakes.

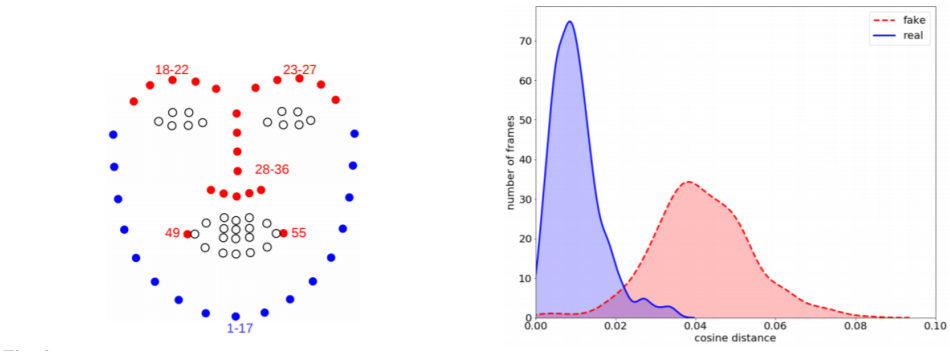


Figure 4: Left: landmarks of facial features, Right: distribution of cosine distance for estimated head pose vectors. [48]

In *Exposing Deep Fakes Using Inconsistent Head Poses* [48], an experiment was conducted where they looked at the head orientation vector in facial images. A rotation matrix was estimated using facial landmarks between real and generated faces and the cosine difference between the two unit vectors were compared. The smaller this value is, the closer the two vectors are to each other, and the results show that the cosine distance of two estimated head pose vectors for the real images are concentrated on a smaller range of values up to 0.02 while images generated by deepfakes range between 0.02 and 0.08. As seen in Figure 4, there is a statistically significant difference between the head pose vectors for real images and those generated by deepfakes.

# 6 Government Hosted Deepfake Detection Platform

Although deepfake detection technology is reasonably accurate for detecting deepfakes and extremely easy to implement; as mentioned in section 5, techniques for identifying deepfakes based on just visual artifacts alone can already achieve accuracy rates of up to 86.6% [4] on videos from the FaceForensics dataset [5]. We propose that with a variety of these media forensics techniques feed into a pipeline, we can build a tool that can identify most deepfakes on the internet.

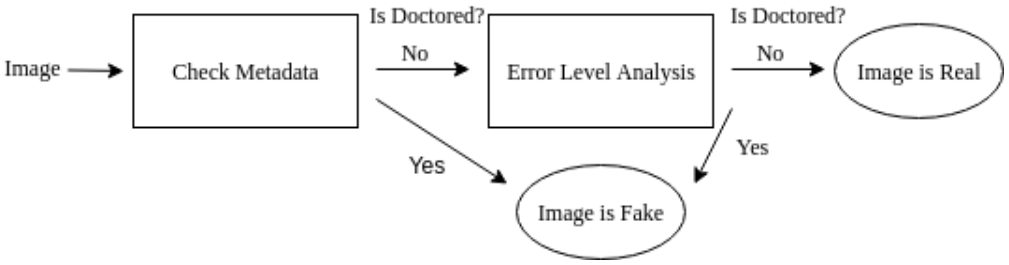


Figure 5: Example of how a pipelined solution is designed containing only metadata and ELA analysis techniques

Because a legal entity can be categorized as either a natural person or a "corporate personhood" [49]. We recommend that in order to serve all members of the public, the government publicly make available two services for each of the two members: 1) an easy-to-use web application for anyone to upload a video, video link, or image for it to be forensically analyzed for legitimacy, and 2) an API that enables any company to detect and label deepfakes uploaded onto their platform. These services should employee the latest deepfake detection technology to ensure it is always updated and reliable.

## 6.1 Deepfake Detection Web Application

A web application (or web app for short) is a computer application that the client can access in a web browser; unlike traditional software you must download and install, web applications can be accessed entirely through the browser. Common web applications are email clients such as Google's Gmail, banking portals, and even online calculators such as Wolfram Alpha.

We recommend that the Federal Trade Commisssion (FTC) host a web application that allows the public to upload images, videos, or video links (of a limited length) for evaluation for probability of legitimacy. In order to preserve anonymity and privacy of information uploaded, no media content should be stored on the government servers, and no identification should be needed to use the website (for example, requiring users to sign up with a user profile that involves inputting their email, legal name, or address). These privacy clauses should further be made transparent to users so that they do not fear the service.

## 6.2 API for Corporations

An application programming interface (API) is an interface or communication protocol that governs the access point for a server, or allows two applications to talk to each other. We specifi-

cally recommend that the government make available a “Representational State Transfer” (REST) web API for corporations to integrate into their applications since most deepfake detection applications are most relevant on social media sites and other online medium.

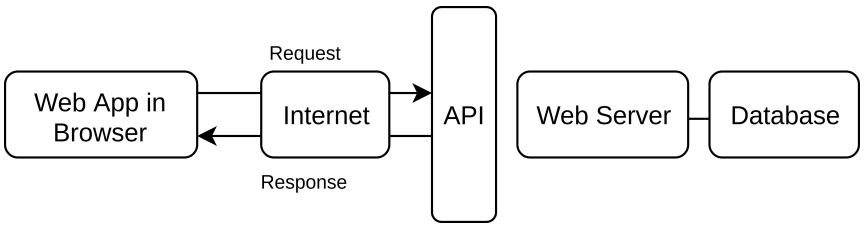


Figure 6: Visual explanation of an API

REST is a standard for how APIs should look like, containing a set of rules for developers to follow. For example, one of the rules states you should get back data called "resource" when you link to a particular URL; each URL is called a request and the data returned is called a response; and each request contains the following four components: endpoint, method, headers, and data (or body).

## 6.3 Recommendation for Open-Sourced Deepfake Detection Platform

One important area of debate is whether or not the code for detecting deepfakes should be open-sourced. We believe that the technology should be open-sourced and anyone should be able to make edit suggestions to it, similar to the structure used during the development of the Linux kernel. The primary reasons for this decision are three-fold: to ensure the technology is always updated and reliable, to maintain public transparency of the algorithms used to detect deepfakes, and to improve cost-effectiveness of hiring a small team for maintenance of the software. The reasons against open-sourcing are that the codebase is subject to manipulation and the systm lacks a proper incentive structure for contributors to the code base. We will cover the pros and argue against the cons in depth in the sections below.

### 6.3.1 Pros for Open-Sourcing

The main argument for why open-sourcing the pipeline for detection of deepfakes helps to ensure technology is always updated and reliable is the same central thesis given in *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary* [6], a book written by Eric S. Raymond on software engineering methods based on his observations of the Linux kernel development process and personal experiences managing fetchmail. Raymond’s central

thesis, which he coined Linus's law, is that "given enough eyeballs, all bugs are shallow". In other words, the more widely available the source code is for public testing, scrutiny, and experimentation, the faster bugs are discovered and progress for software development can be made.

In our case, we want the deep fake technology to always be updated. It is very difficult for a government-employed team to properly update the technology to guard against new innovations in deepfakes at all times because this would involve hiring a research team on the same tier as a research institute. By open-sourcing the platform, the bugs mentioned in Raymond's book, which are analogous to loopholes for which deepfakes are not detected by our platform, can be patched very quickly by volunteers. Due to the services for deepfake detection benefiting both individuals and corporations, some form of corporate sponsorship in the form of monetary compensation or prestige can be granted to those who make substantial contributions to the deepfake detection platform.

Further, open-sourcing the deepfake detection platform enables public transparency of the algorithms used. In modern times when algorithms make many of our decisions for us, it is of critical importance that the algorithms do not harbor biases such as the one unearthed by the Wall Street Journal in 2010 that revealed minorities were directed to apply for cards with higher interest rates than those directed to white visitors on the Capital One banking site [50]. According to Cynthia Dwork of Microsoft Research, quoted at the *Fairness, Accountability, and Transparency in Machine Learning* conference in 2014: "What we advocate is sunshine for the metric. The metric should at the very least be open and up for discussion. There should not be secret metrics." [51]

Finally, open-sourcing the deepfake detection platform means that the government does not have to hire an expensive team to maintain and provide all technical updates to the code. Not only is this very expensive, but there is no guarantee that this team is a fail-proof guard against missing bugs or failing to observe particular technical trends, as mentioned above. By open-sourcing the code after initial development, the government only needs to hire a small team responsible for reviewing code submitted by community members in order to decide whether to reject or approve the change; this team would be solely responsible for maintaining the codebase as opposed to actively pushing it forward technically.

### 6.3.2 Rebutting Arguments Against Open-Sourcing

The two main arguments against open-sourcing the codebase for deepfake detection is that the codebase could be subject to manipulation and there is a lack of proper incentive structure for community members to make contributions. We will address both of these concerns below.

In response to the first concern that the codebase could be subject to manipulation, there are two barriers preventing malicious actors from making permanent changes. The first is that a small team is responsible for reviewing and approving code pushes. Second, due to the code base's open-source nature, other community contributors that notice the approved manipulation can submit code with comments to undo or fix the manipulation.

Second, a lack of proper incentive structure for community members to make contributions can be overturned by cash prizes and prestige awarded to substantial contributors. Because the deep fake detector API offers substantial utility to both companies and individuals, there could a donation or sponsorship in place to support monetary awards.

| Pros | Cons |
|---|---|
| - **Cost-effectiveness**: | - **Subject to manipulation**: |
| The government does not need to hire a team to make all technical updates to the deepfake detection platform. Only a small team is necessary to approve push requests for code to the repository submitted by individual contributors unafilliated by the government. | A malicious actor could submit code containing loopholes for particular flavors of deepfakes. However, the small team employed to review source code is responsible for reading requests careful and can deny the push request. Further, others can notice manipulation and push code to fix it. |
| - **Fast iteration**: | - **Lack of proper incentive strucuture:** |
| Rapid development of technical improvements and discovery of bugs. Due to the "given enough eyeballs, all bugs are shallow" thesis given in *The Cathedral and the Bazaar* [6]. | There is no reason for people to maintain the code-base for a government sponsored deepfake detection platform. However, because the service releases a free API for corporations, corporations are incentivized to monetarily sponsor the service. Prestigious recognition and possible monetary prizes can be awarded to active and subtantial contributors. |
| - **Algorithmic transparency**: | |
| To ensure that algorithms are fair and unbiased, maintaining algorithmic transparency allows the technical public to review source code and evaluate the algorithms used. | |

Table 1: Summary of pros and cons for open-sourcing deepfake detection platform

# 7 Conclusion

Deepfakes are becoming increasingly more realistic and can be generated very quickly at scale thanks to apps such as *FakeApp* that allow users to make deepfaked faceswaps without any technical background. Because of the high concentration of users around social media platforms and the blurred lines between what is fake and real due to deepfake's realism, it is easy to spread false information so that deepfakes pose a serious threat to democracy, national security, and public

policy. In section 3, we have produced a variety of both past incidents and hypothetical future scenarios that emphasize the danger of deepfakes, as well as analysis of analogous judicial precedence that should guide us in interpreting deepfake cases in Court in section 4.

To guard against these dangers, we recommend that the government release a deepfake detection platform using a pipeline of forensic tools that is available to the public in two forms: a web application with an easy-to-use user interface that allows anyone to upload a image or video to be analyzed, and a REST API for corporations to use in their applications. To support this development, we have proposed a list of technical options for detecting deepfakes that include metadata filters, error level analyses, visual artifacts, and inconsistent head poses. However, we recognize that the tools for detecting deepfakes are ever evolving in tandum with the evolution of deepfakes. Therefore, we further recommend that in order for this government-hosted deep fake detection tool to be up to date, cost-effective, and algorithmically transparent, the government should open source the code base and encourage community engagement from the technical open-source community.

# 8 Acknowledgements

We would like to thank Zachary Pitcher and Anish Athalye for thoughtful discussions. We would like to thank Zachary Pitcher again for providing emotional support throughout this project. Also we would like to thank the 6.805 instructors for being understanding and supportive, especially to Michael Trice and David Edelman for their feedback and kindness.

# 9 Division of Work

Catherine: Section 2: What are Deepfakes, Section 5: Techniques for Identifying Deepfakes, Section 6: Government Hosted Deepfake Detection Platform (technical), Executive Summary, Conclusion.

Rafael: Section 3: Threats to Democracy, National Security, and Public Policy, Section 4: Analogous Judicial Precedent, Section 6: Government Hosted Deepfake Detection Platform (policy), Executive Summary, Conclusion.

# References


[1] R. Cellan-Jones, "Deepfake videos 'double in nine months'," Oct 2019. [Online]. Available: https://www.bbc.com/news/technology-49961089

[2] "Creating a data set and a challenge for deepfakes." [Online]. Available: https://ai.facebook.com/blog/deepfake-detection-challenge/

[3] A. Breland, "The bizarre and terrifying case of the "deefake" video that helped bring an african nation to the brink," March 2019. [Online]. Available: https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/

[4] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 2019.

[5] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics: A large-scale video dataset for forgery detection in human faces," *ArXiv*, vol. abs/1803.09179, 2018.

[6] E. S. Raymond, *Cathedral and the bazaar*. SnowBall Publishing, 2010.

[7] M. Schreyer, T. Sattarov, B. Reimer, and D. Borth, "Adversarial learning of deepfakes in accounting." [Online]. Available: https://arxiv.org/pdf/1910.03810.pdf

[8] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," 12 2018.

[9] "Sham00k." [Online]. Available: https://www.youtube.com/channel/UCZXbWcv7fSZFTAZV4beckyw

[10] J. Meskimen, "A deeper look into the life of an impressionist," Oct 2019. [Online]. Available: https://www.youtube.com/watch?v=5rPKeUXjEvE

[11] "Top 10 doctored photos - photo essays." [Online]. Available: http://content.time.com/time/photogallery/0,29307,1924226_1949526,00.html

[12] S. Greenwood, A. Perrin, and M. Duggan, "Demographics of social media users in 2016," May 2017. [Online]. Available: https://www.pewresearch.org/internet/2016/11/11/social-media-update-2016/

[13] D. Harwell, "White house shares doctored video to support punishment of journalist jim acosta," Nov 2018. [Online]. Available: https://www.washingtonpost.com/technology/2018/11/08/white-house-shares-doctored-video-support-punishment-journalist-jim-acosta/

[14] J. Amatulli, "People on twitter call for sarah huckabee sanders to resign for jim acosta video," Nov 2018. [Online]. Available: https://www.huffpost.com/entry/people-call-for-sarah-huckabee-sanders-to-resign-for-jim-acosta-video_n_5be448e6e4b0dbe871a8109d

[15] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[16] "Astronomers explore uses for ai-generated images." [Online]. Available: https://www.nature.com/news/astronomers-explore-uses-for-ai-generated-images-1.21398

[17] L. Matney, "Scale ai and its 22-year-old ceo lock down $100 million to label silicon valley's data," Aug 2019. [Online]. Available: https://techcrunch.com/2019/08/05/scale-ai-and-its-22-year-old-ceo-lock-down-100-million-to-help-label-silicon-valleys-data/

[18] "List of datasets for machine-learning research," Nov 2019. [Online]. Available: https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research

[19] "The generator | generative adversarial networks | google developers." [Online]. Available: https://developers.google.com/machine-learning/gan/generator

[20] J. Grubb, "Machine learning is rescuing old game textures in zelda and final fantasy," Jan 2019. [Online]. Available: https://venturebeat.com/2019/01/18/machine-learning-is-rescuing-old-game-textures-in-zelda-and-final-fantasy/

[21] K. Schawinski, C. Zhang, H. Zhang, L. Fowler, and G. K. Santhanam, "Generative adversarial networks recover features in astrophysical images of galaxies beyond the deconvolution limit," *Monthly Notices of the Royal Astronomical Society: Letters*, 2017.

[22] K. Poulsen, "We found the guy behind the viral 'drunk pelosi' video," June 2019. [Online]. Available: https://www.thedailybeast.com/we-found-shawn-brooks-the-guy-behind-the-viral-drunk-pelosi-video

[23] D. Trump, "Donald trump tweet of doctored pelosi video," May 2019. [Online]. Available: https://twitter.com/realDonaldTrump/status/1131728912835383300?s=20

[24] A. O'Reilly, "Biden campaign accused of deceptive editing in new anti-trump ad," December 2019. [Online]. Available: https://www.foxnews.com/politics/biden-campaign-accused-of-deceptive-editing-in-new-anti-trump-ad

[25] M. Jacobs, "Osama bin laden tape." [Online]. Available: https://cphcmp.smu.edu/2004election/osama-bin-laden-tape/

[26] P. Harris, P. Beaumont, and J. Burke, "Bush wins boost from terror tape," October 2004. [Online]. Available: https://www.theguardian.com/world/2004/oct/31/alqaida.uselections20041

[27] D. K. Citron and R. Chesney, "Deep fakes: A looming crisis for national security, democracy and privacy?" [Online]. Available: https://scholarship.law.bu.edu/shorter_works/33

[28] L. A. Stanton, "Donald trump claims 2005 'access hollywood' tape is not authentic: Report"," November 2017. [Online]. Available: https://www.usmagazine.com/celebrity-news/news/donald-trump-claims-2005-access-hollywood-tape-is-fake/

[29] "President ali bongo ondimba delivers new year's address video," January 2019. [Online]. Available: https://www.facebook.com/tvgabon24/videos/324528215059254/?v=324528215059254

[30] C. McGreal, "Wikileaks reveals video showing us air crew shooting down iraqi civilians," April 2010. [Online]. Available: https://www.theguardian.com/world/2010/apr/05/wikileaks-us-army-iraq-attack

[31] October 2019. [Online]. Available:
https://www.archives.gov/education/lessons/zimmermann

[32] "Report on the investigation into russian interference in the 2016 presidential election," May 2019. [Online]. Available: https://www.justice.gov/storage/report.pdf

[33] D. Lee, "The tactics of a russian troll farm," February 2018. [Online]. Available:
https://www.bbc.com/news/technology-43093390

[34] ——, "Deepfake porn has serious consequences," February 2018. [Online]. Available:
https://www.bbc.com/news/technology-42912529

[35] R. Hasen, "Deep fakes, bots, and siloed justices: American election law in a post-truth world," *St. Louis University Law Journal*. [Online]. Available:
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3418427

[36] H. Norton, "Lies and the constitution," *The Supreme Court Review*, vol. 2012, no. 1, p. 161–201, 2013.

[37] "Deepfakes: False pornography is here and the law cannot protect you," Jan 2019. [Online]. Available: https://dltr.law.duke.edu/2019/01/05/
deepfakes-false-pornography-is-here-and-the-law-cannot-protect-you/

[38] E. Gibbs, "China seeks to root out fake news and deepfakes with new online content rules," November 2019. [Online]. Available:
https://www.reuters.com/article/us-china-technology/
china-seeks-to-root-out-fake-news-and-deepfakes-with-new-online-content-rules-idUSKBN1Y30VU

[39] "Defamation." [Online]. Available: https://www.law.cornell.edu/wex/defamation

[40] "Intentional infliction of emotional distress." [Online]. Available:
https://www.law.cornell.edu/wex/intentional_infliction_of_emotional_distress

[41] D. P. Sacks, "Snyder v. phelps, the supreme court's speech-tort jurisprudence, and normative considerations." [Online]. Available: https://www.yalelawjournal.org/forum/
snyder-v-phelps-the-supreme-courts-speech-tort-jurisprudence-and-normative-considerations

[42] "Using the name or likeness of another." [Online]. Available:
http://www.dmlp.org/legal-guide/using-name-or-likeness-another

[43] E. P. Robinson, "False light." [Online]. Available:
https://www.mtsu.edu/first-amendment/article/957/false-light

[44] "False light." [Online]. Available: http://www.dmlp.org/legal-guide/false-light

[45] A. Hill, "Defamation and privacy under the first amendment," *Columbia Law Review*, vol. 76, no. 8, p. 1205, 1976.

[46] S. Tariq, S. Lee, H. Kim, Y. Shin, and S. S. Woo, "Detecting both machine and human created fake face images in the wild," *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security - MPS 18*, 2018.

[47] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niebner, "Face2face: Real-time face capture and reenactment of rgb videos," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[48] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[49] S. Pruitt, "How the 14th amendment made corporations into 'people'," Jun 2018. [Online]. Available: https://www.history.com/news/14th-amendment-corporate-personhood-made-corporations-into-people

[50] E. Steel and J. Angwin, "On the web's cutting edge, anonymity in name only," Aug 2010. [Online]. Available: https://www.wsj.com/articles/SB10001424052748703294904575385532109190198

[51] "Racist in the machine: The disturbing implications of algorithmic bias," *Human Rights Documents Online*.