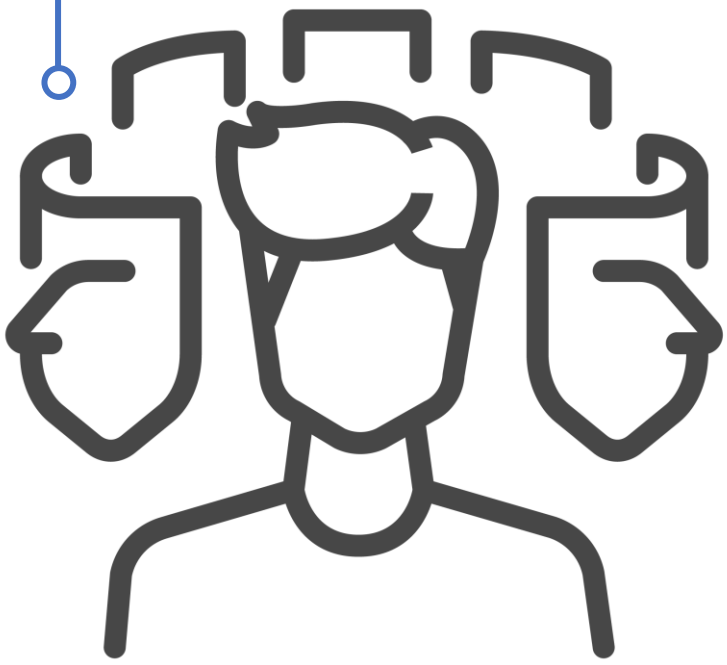
Increasing Threat of

DEEPFAKE

Identities



DISCLAIMER STATEMENT: *This document is provided for educational and informational purposes only. The views and opinions expressed in this document do not necessarily state or reflect those of the United States Government or the Public-Private Analytic Exchange Program, and they may not be used for advertising or product endorsement purposes. All judgments and assessments are solely based on unclassified sources and are the product of joint public and private sector efforts.*

Increasing Threats of Deepfake Identities

Abstract

Deepfakes, an emergent type of threat falling under the greater and more pervasive umbrella of synthetic media, utilize a form of artificial intelligence/machine learning (AI/ML) to create believable, realistic videos, pictures, audio, and text of events which never happened.

Many applications of synthetic media represent innocent forms of entertainment, but others carry risk.

The threat of Deepfakes and synthetic media comes not from the technology used to create it, but from people's natural inclination to believe what they see, and as a result deepfakes and synthetic media do not need to be particularly advanced or believable in order to be effective in spreading mis/disinformation.

Based on numerous interviews conducted with experts in the field, it is apparent that the severity and urgency of the current threat from synthetic media depends on the exposure, perspective, and position of who you ask. The spectrum of concerns ranged from "an urgent threat" to "don't panic, just be prepared."

To help customers understand how a potential threat might arise, and what that threat might look like, we considered a number of scenarios specific to the arenas of commerce, society, and national security.

The likelihood of any one of these scenarios occurring and succeeding will undoubtedly increase as the cost and other resources needed to produce usable deepfakes simultaneously decreases - just as synthetic media became easier to create as non-AI/ML techniques became more readily available.

In line with the multifaceted nature of the problem, there is no one single or universal solution, though elements of technological innovation, education, and regulation must comprise part of any detection and mitigation measures.

In order to have success there will have to be significant cooperation among stakeholders in the private and public sectors to overcome current obstacles such as "stovepiping" and to ultimately protect ourselves from these emerging threats while protecting civil liberties.

TEAM INTRODUCTIONS

Tina Brooks, Verizon
Princess G., Transportation Security Administration
Jesse Heatley, JP Morgan Chase & Co.
Jeremy J., United States Secret Service
Scott Kim, Experian
Samantha M., Federal Bureau of Investigation
Sara Parks, National Cyber-Forensics & Training Alliance
Maureen Reardon, Melian LLC
Harley Rohrbacher, National Cyber-Forensics & Training Alliance
Burak Sahin, Deloitte & Touche
Shani S., Federal Bureau of Investigation (Co-Champion)
James S., U.S. Department of Homeland Security
Oliver T., Federal Bureau of Investigation
Richard V., Federal Bureau of Investigation (Co-Champion)

TERMINOLOGY ACKNOWLEDGEMENTS: The terms "Kleenex," "Xerox," and "Photoshop" once represented specific products from a single manufacturer, yet today, through common use (or mis-use) they are universally recognized as representative of a class of products, regardless of manufacturer. Across the broad population, the term "deepfakes" appears to have acquired a similar connotation to any synthetic media. Our team does not endorse such mis-use of the term, but we are pragmatists. As a result, in this paper we will occasionally use the term "deepfakes" to refer to any type of synthetic media, regardless of whether it truly represents a "deepfake."

Likewise, in this paper the terms Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) are frequently mentioned in relation to deepfakes and other synthetic media. As noted elsewhere, both machine learning and deep learning may be considered as subsets of artificial intelligence enabling techniques. Therefore, in some cases we may use the term "AI/ML" where some experts might have preferred the use of "ML" or "DL."

Introduction

In late 2017, Motherboard reported on a video that had appeared on the Internet in which the face of Gal Gadot had been superimposed on an existing pornographic video to make it appear that the actress was engaged in the acts depicted.[1] Despite being a fake, the video quality was good enough that a casual viewer might be convinced – or might not care.

An anonymous user of the social media platform Reddit, who referred to himself as "deepfakes," claimed to be the creator of this video."[2]

The term "deepfakes" is derived from the fact that the technology involved in creating this particular style of manipulated content (or "fakes") involves the use of deep learning techniques. Deep learning represents a subset of machine learning techniques which are themselves a subset of artificial intelligence. In machine learning, a model uses training data to develop a model for a specific task. The more robust and complete the training data, the better the model gets. In deep learning, a model is able to automatically discover representations of features in the data that permit classification or parsing of the data. They are effectively trained at a "deeper" level.[3]

The data which can be examined using deep learning is not restricted to images and videos of people. It can include images and videos of anything, as well as audio and text. In 2020, Dave Gershgorn, a reporter for OneZero reported on the release of "new" music by famous artists on the OpenAI website.[4] Using existing tracks from well-known artists, living and dead, programmers were able to create realistic tracks of new songs by Elvis, Frank Sinatra, and Jay-Z. Jay-Z's company, Roc Nation LLC, sued YouTube to take the tracks down.[5]

AI-generated text is another type of deepfake that is a growing challenge. Whereas researchers have identified a number of weaknesses in image, video, and audio deepfakes as means of detecting them, deepfake text is not so easy to detect.[6] It is not out of the question that a user's texting style, which can often be informal, could be replicated using deepfake technology.

All of these types of deepfake media – image, video, audio, and text – could be used to simulate or alter a specific individual or the representation of that individual. This is the primary threat of deepfakes. However, this threat is not restricted to deepfakes alone, but incorporates the entire field of "Synthetic Media" and their use in disinformation.

More than just "deepfakes" – "Synthetic Media" and Disinformation

Deepfakes actually represent a subset of the general category of "synthetic media" or "synthetic content." Many popular articles on the subject[7][8] define synthetic media as any media which has been created or modified through the use of artificial intelligence/machine learning (AI/ML), especially if done in an automated fashion. From a practical standpoint, however, within the law enforcement and intelligence communities, synthetic media is generally defined to encompass all media which has either been created through digital or artificial means (think computer-generated people) or media which has been modified or otherwise manipulated through the use of technology, whether analog or digital. For

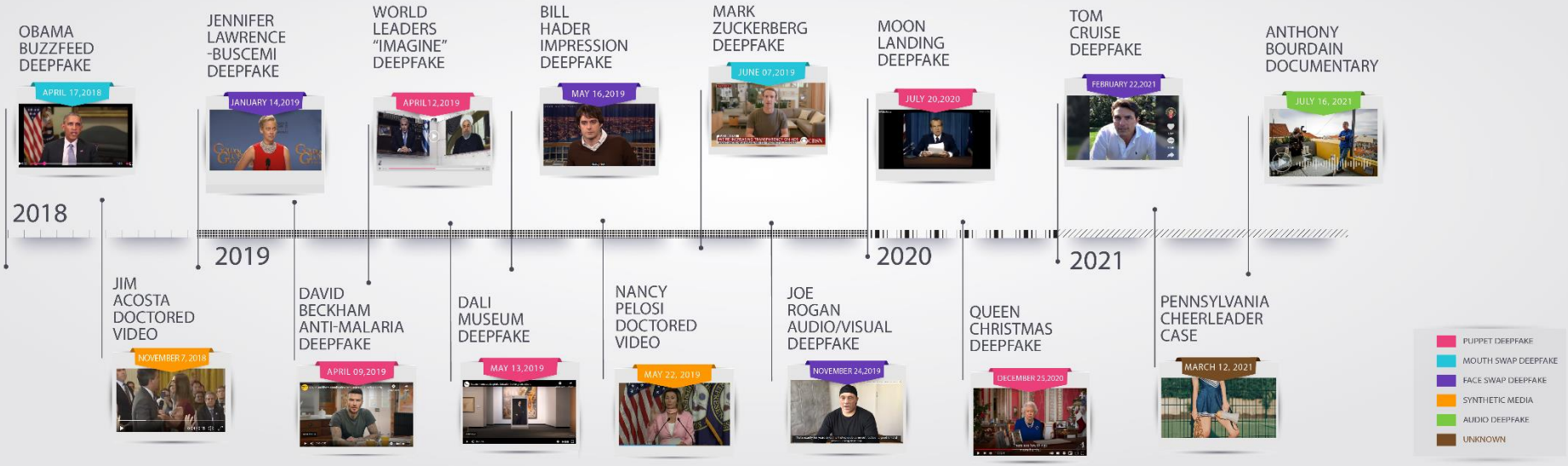
example, physical audio tape can be manually cut and spliced to remove words or sentences and alter the overall meaning of a recording's content. "Cheapfakes" are another version of synthetic media in which simple digital techniques are applied to content to alter the observer's perception of an event. Cheapfake examples described elsewhere in this paper demonstrate speech being slowed, and video being accelerated.

Science and technology are constantly advancing. Deepfakes, along with automated content creation and modification techniques, merely represent the latest mechanisms developed to alter or create visual, audio, and text content. The key difference they represent, however, is the ease with which they can be made – and made well. In the past, casual viewers (or listeners) could easily detect fraudulent content. This may no longer always be the case and may allow any adversary interested in sowing misinformation or disinformation to leverage far more realistic image, video, audio, and text content in their campaigns than ever before.

How are deepfakes made and how might they be used?

Since the first deepfake in 2017, there have been many developments in deepfake and related-synthetic media technologies. The timeline below provides a listing of some of the most well-known and representative examples of deepfakes, as well as some "cheapfakes" and one example of an instance in which deepfakes were initially implicated, but never proven to have been used. An addendum to this report, which provides summaries of these examples and links for further information is also available.

TIMELINE OF DEEPFAKES AND SYNTHETIC MEDIA



Deepfake and Synthetic Media Example Links for More Information

Obama Buzzfeed: https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed

Jim Acosta Doctored Video: https://apnews.com/article/entertainment-north-america-donald-trump-us-news-ap-top-news-c575bd1cc3b1456cb3057ef670c7fe2a

Jennifer Lawrence- Steve Buscemi: https://fortune.com/2019/01/31/what-is-deep-fake-video/

David Beckham Anti-Malaria PSA: https://www.campaignlive.com/article/deepfake-voice-tech-used-good-david-beckham-malaria-campaign/1581378

World Leaders Sing "Imagine": https://scifi.radio/2019/05/29/watch-world-leaders-sing-for-peace-in-canny-ais-imagine-video/

Dali Museum: https://www.dezeen.com/2019/05/24/salvador-dali-deepfake-dali-musuem-florida/

Bill Hader Impressions: https://www.fastcompany.com/90353902/bill-haders-al-pacino-impression-gets-even-more-real-and-creepy-with-the-help-of-deepfakes

Nancy Pelosi Doctored Video: https://www.usatoday.com/story/news/factcheck/2020/08/11/fact-check-video-pelosi-altered-and-selectively-edited/3332920001/

Mark Zuckerberg: https://www.technologyreview.com/2019/06/12/134992/facebook-deepfake-zuckerberg-instagram-social-media-election-video/

Joe Rogan: https://www.maxim.com/news/joe-rogan-audio-and-video-deepfake-2019-12

Nixon/Moon Landing: https://www.newsweek.com/richard-nixon-deepfake-apollo-disinformation-mit-1475340

Queen's Christmas Speech: https://www.independent.co.uk/news/uk/home-news/queen-deepfake-channel-4-christmas-message-b1778542.html

Tom Cruise TikToks: : https://www.theverge.com/22303756/tiktok-tom-cruise-impersonator-deepfake

Pennsylvania Cheerleader Case: https://www.buckscountycouriertimes.com/story/news/2021/05/14/da-chalfont-woman-may-not-have-used-deepfake-tech-harassment-vipers-cheerleading/4992798001/

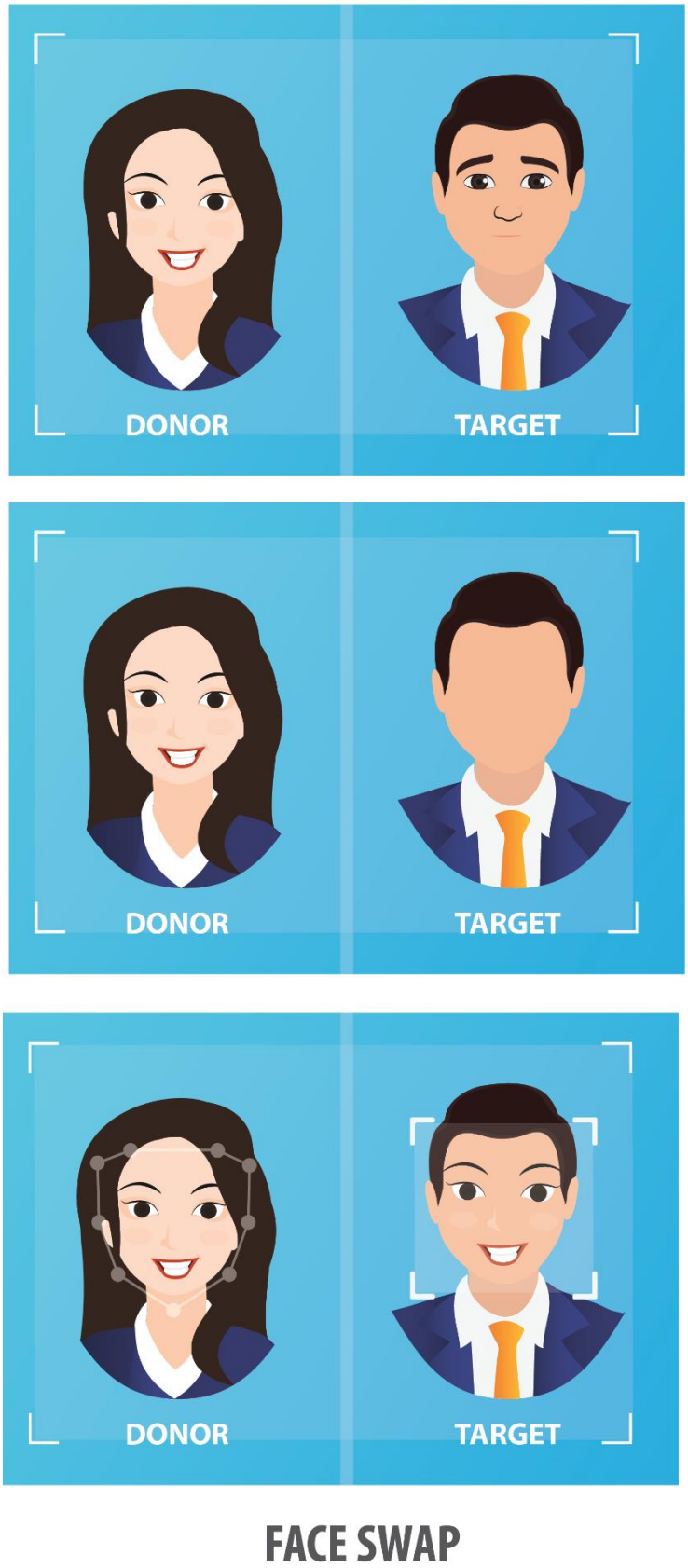
Anthony Bourdain Documentary: https://www.newyorker.com/culture/annals-of-gastronomy/the-ethics-of-a-deepfake-anthony-bourdain-voice

The most common techniques for creating deepfakes are represented on the timeline. The first type, which pre-dates deepfake and AI/ML technology, is the **face swap**.

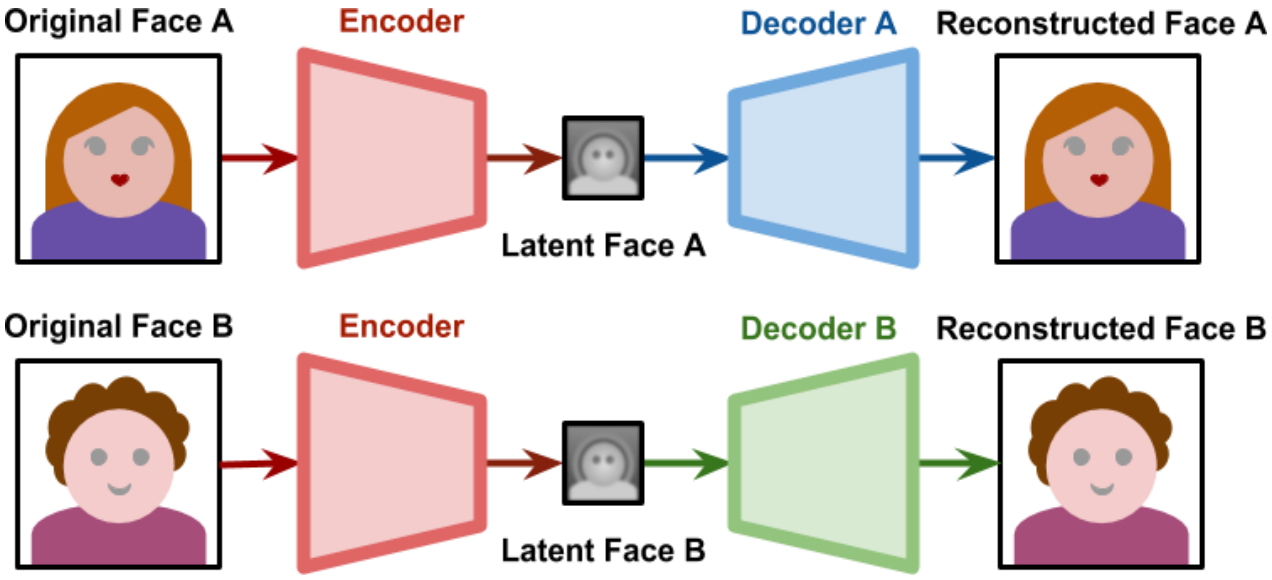
In the 1990's, with commercial distribution of image editing software, such as Adobe© Photoshop™, it became possible for anyone with a computer to alter an image, including putting the face or head of one person onto the body of another person.

Today, the technology utilized to produce a convincing face swap involves AI. The technology allows an adversary to swap the face of one person onto another person's face and body. The adversary could use Encoder or Deep Neural Network (DNN) technology to create a face swap. To learn the face swap model using an autoencoder, preprocessed samples of person A and person B are mapped to the same intermediate compressed latent space using the same encoder parameters[9].

Once the three networks are trained, to swap the face of person B onto A, the target video (or image) of A is fed frame by frame into the common encoder network, and then decoded by person B's decoder network[10]. There are many applications that allow a user to swap faces but not all use the same technology. Some of the applications that allow a user to perform a face swap are FaceShifter, FaceSwap, DeepFace Lab, Reface, and TikTok. Apps like Snapchat and TikTok offer dramatically lowered computational and expertise requirements that allow users to generate various manipulations in real time[11].



FACE SWAP



When people think of a deepfake face swap, they commonly think of a video like the one of Bill Hader transforming into Tom Cruise, Arnold Schwarzenegger, Al Pacino and Seth Rogan on David Letterman's show in 2019. This is an example of deepfake use that is not harmful. But as we've seen, there is a dark side to AI/ML and it has nothing to do with the technology, but with the people using it.

The world needs to be aware of the inherent risk of deepfakes by malign actors.

One major harmful use of face swap technology is deepfake pornography. Face swap technology was used to put actors Kristin Bell and Scarlett Johansson in several pornographic videos. One of the fake videos that was labeled as "leaked" footage generated over 1.5 million views.[13] Women have no way of preventing a malign actor from creating deepfake pornography.

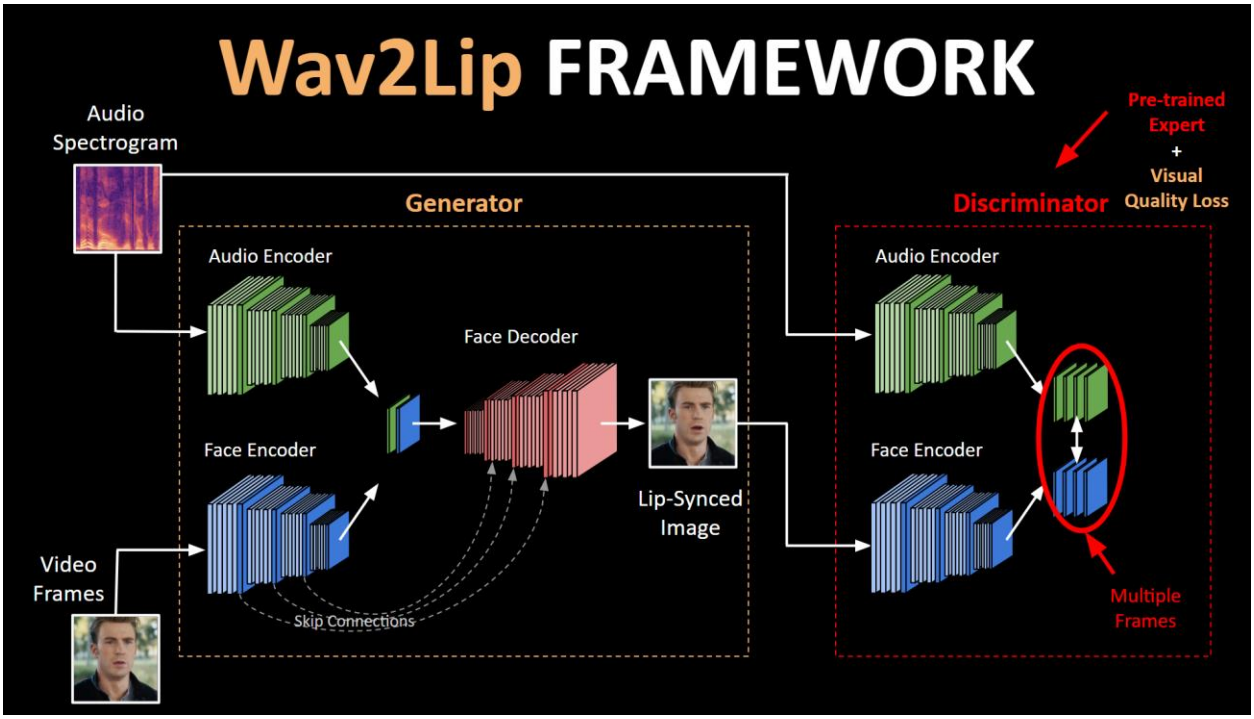
The use of the technology to harass or harm private individuals who do not command public attention and cannot command resources necessary to refute falsehoods should be concerning[14]. The ramifications of deepfake pornography have only begun to be seen.

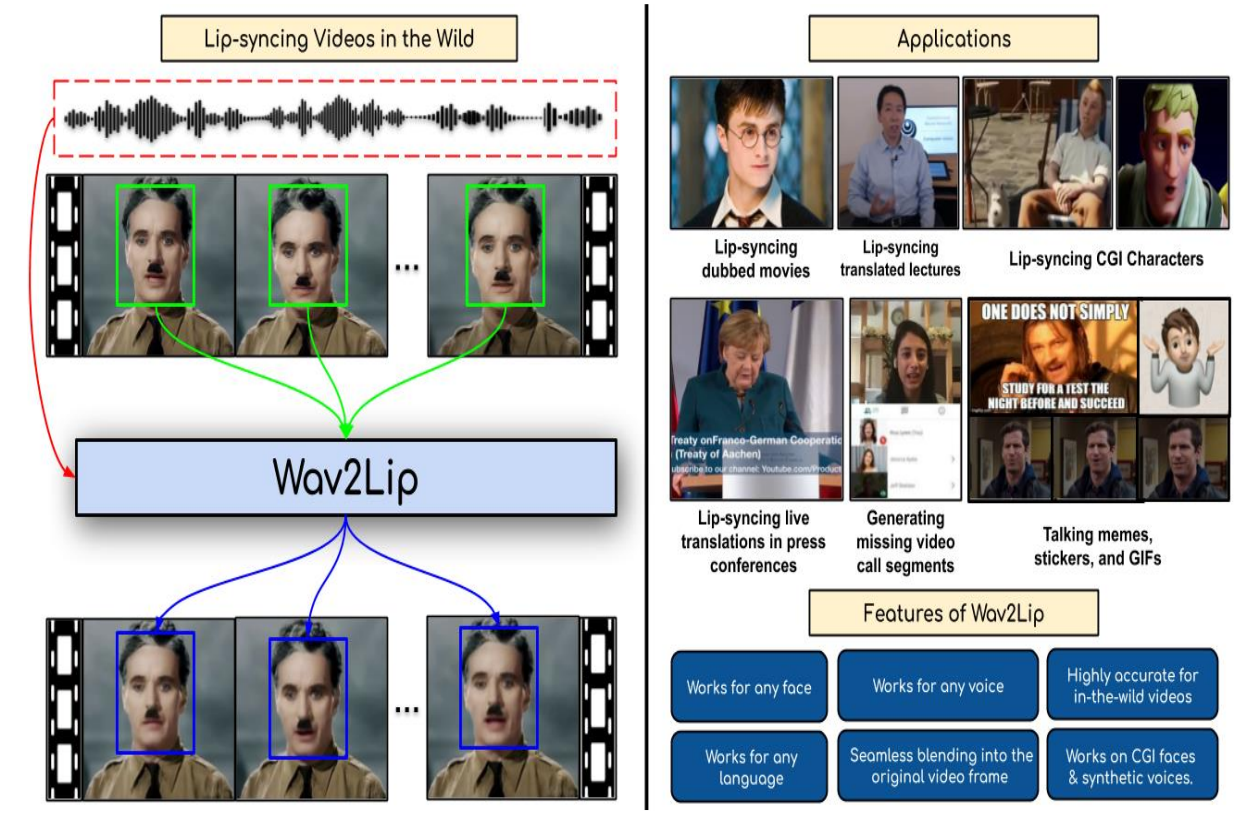
Another deepfakes technique is "**Lip Syncing**." Lip Syncing involves "mapping [a] voice recording from one or multiple contexts to a video recording in another, to make the subject of the video appear to say something authentic[15]." Lip synching technology allows the user to make their target say anything they want through the use of recurrent neural networks (RNN). In November 2017 Stanford researchers published a paper and model for Face2Face, a RNN based video production model that allows third parties the ability to put words in the mouth of public figures in real time[16].

Deepfake technology has grown rapidly since then and has become more available to the general population. As these techniques become more accessible the risk of harm to private

figures will increase, especially those who are politically, socially, or economically vulnerable[17].

A new AI/ML technology called Wav2Lip has enabled the creation of a lip synch deepfake. [18] [19] .





Wav2Lip produces extremely realistic audio and is the first speaker-independent model to generate videos with lip-sync accuracy that matches the real synced videos[20]. Human evaluations indicate that the generated videos of Wav2Lip are preferred over existing methods and unsynced versions more than 90% of the time

A final deepfake technique represented on the timeline is referred to as the '**puppet**' technique. As the name implies, the puppet technique allows the user to make the targeted individual move in ways they did not actually move. This can include facial movements or whole body movements. Puppet deepfakes use Generative Adversarial Network (GAN) technology that consists of computer-based graphics. Refer to the text box below for more about GANs.

Generative Adversarial Networks (GANs)

A key technology leveraged to produce deepfakes and other synthetic media is the concept of a "Generative Adversarial Network" or GAN. In a GAN, two machine learning networks are utilized to develop synthetic content through an adversarial process. The first network is the "generator." Data that represents the type of content to be created is fed to this first network so that it can 'learn' the characteristics of that type of data. The generator then attempts to create new examples of that data which exhibit the same characteristics of the original data. These generated examples are then presented to the second machine learning network, which has also been trained (but through a slightly different approach) to 'learn' to identify the characteristics of that type of data. This second network (the "adversary") attempts to detect flaws in the presented examples and rejects those which it determines do not exhibit the same sort of characteristics as the original data – identifying them as "fakes." These fakes are then 'returned' to the first network, so it can learn to improve its process of creating new data. This back and forth continues until the generator produces fake content that the adversary identifies as real. The first practical application of GANs was established by Ian Goodfellow and his coworkers in 2014, when they demonstrated the ability to create synthetic images of human faces.[1] While human faces are a popular subject of GANs, they can be applied to any content. The more detailed (i.e., realistic) the content used to train the networks in a GAN, the more realistic the output will be.

FULL PUPPET

Unlike the learning-based methods, some methods rely on more traditional computer-graphics approaches to create deepfakes. Face2Face, for example, allows for the creation of so-called puppet master deepfakes in which one person's (the master's) facial expression and head movements are mapped onto another person (the puppet)[21].

Deepfakes In Relation to Cheapfakes and the State of the Art in Digital Manipulation

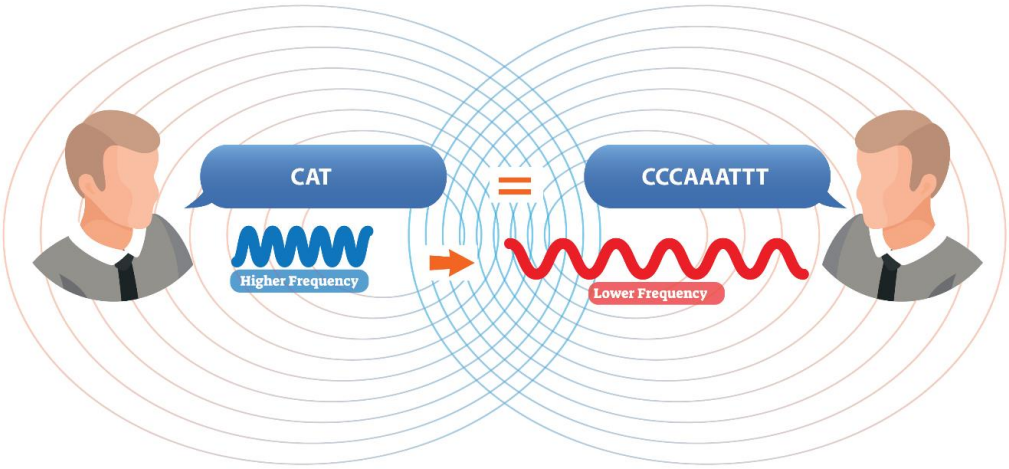
Society has been able to manipulate synthetic media for centuries. Audio-visual cheap fakes even pre-date the digital age.

In the lead-up to UK elections in 1983, members of the British anarcho-punk band Crass spliced together excerpts from speeches by Margaret Thatcher and Ronald Reagan to create a fake telephone conversation between the leaders, in which they each made bellicose, politically damaging statements[22].

Before there was AI/ML there was Adobe© Photoshop™, slowing down or speeding up of video, and lookalikes.

Today, these techniques are called "Cheapfakes" and are also known as "shallow fakes." These are audiovisual (AV) manipulations created with cheaper, more accessible software (or none at all) as compared to deepfakes[23]. These techniques are less expensive, require

less technical skill, and are available on a larger scale. One of the most famous recent cheapfake videos, depicted Speaker of the House Nancy Pelosi in a video in which her speech was slowed down to make her appear intoxicated. When deepfakes were first developed several years ago, their creation required a high level of skill in AI, training, and technology, along with advanced equipment, training data and other resources such as time. Recent developments, however, have now made deepfake technology far less resource intensive, and so more accessible to the general population.

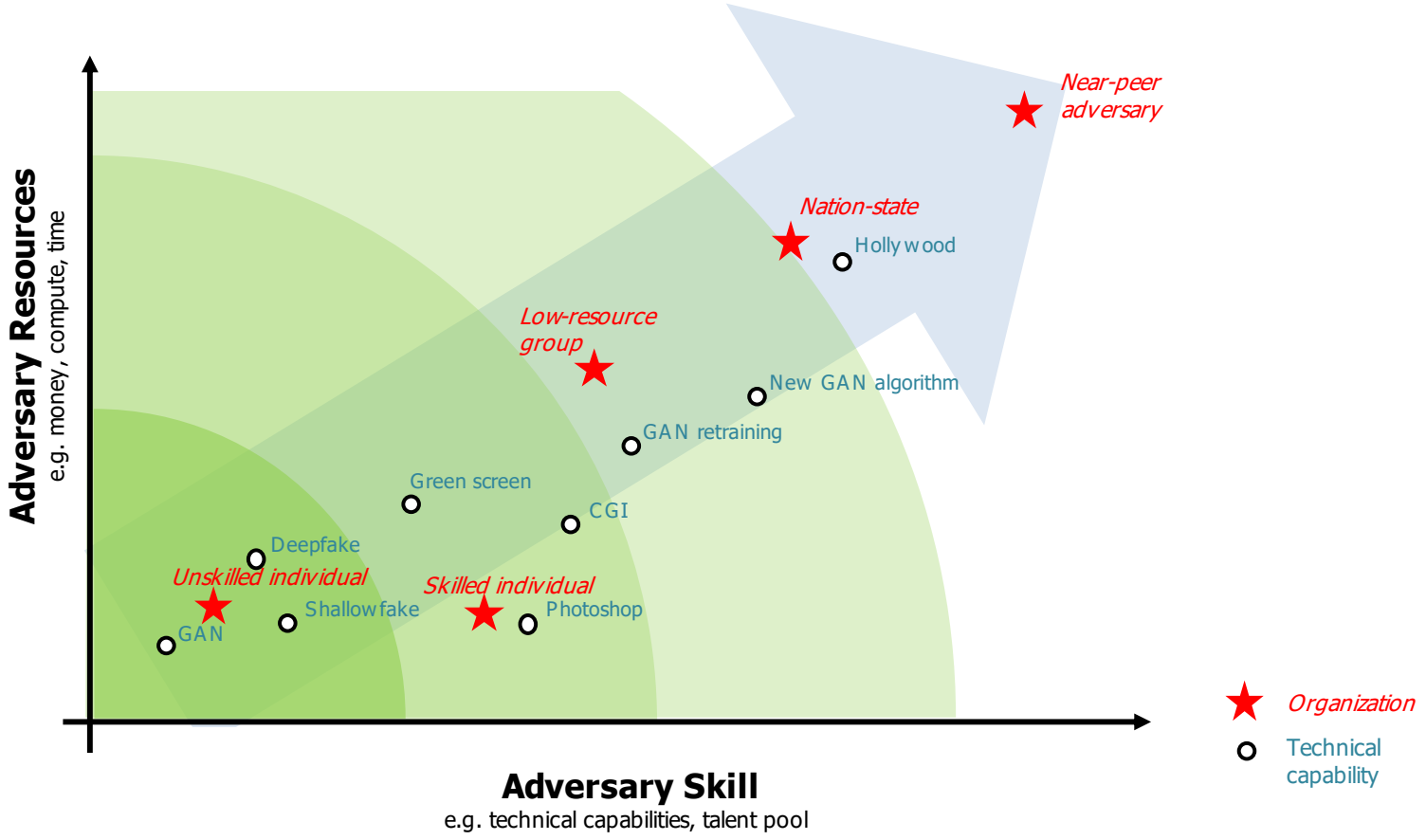


AUDIO/ VIDEO SPEED REDUCTION

In 2015, the Defense Advanced Research Projects Agency (DARPA) launched the Media Forensics, or "MediFor" program. MediFor aimed to automatically detect manipulations in visual media (images and videos), provide detailed information about the manipulation, or manipulations, detected, and reason about the overall integrity of the media to help end users authenticate any questionable image or video.

To place MediFor developments within the broader context of synthetic media and techniques for content manipulation, the MediFor Program mapped out the state of the art in synthetic media and visual manipulation technologies. The chart below is designed to map different technologies against the level of skill (horizontal axis) and level of resources (i.e., time, money and compute power – vertical axis) needed to utilize the different technique or technology.

Adversarial landscape



As noted in the text box above, Generative Adversarial Networks (GANs) represent a key evolution in the creation of realistic synthetic content. There is not a single GAN, which is why the chart above shows multiple GAN entries, including the possibility of new ones. The chart shows that deepfakes today require a low level of skill and resources.

Any form of synthetic media manipulation can be weaponized to cause damage. The rapid advancement of deepfakes, however, increases the threat. Deepfakes are more realistic than Cheapfakes and harder to detect. Dr. Matthew Wright, Director of Research for the Global Cybersecurity Institute and a Professor of Computing Security at the Rochester Institute of Technology, told the AEP Team that '…cheapfakes must be easier to do because we see so much more of them.' However, deepfakes are the emerging, weightier issue. It is the difference between a calculator and a computer; the computer is exponentially more powerful and capable of doing all the things; more powerful for what you can do; more visually compelling, accurate, and higher resolution; anything you can do with cheap fakes you can do with deepfakes but more believable .

THREAT ENVIRONMENT

Deepfakes and the misuse of synthetic content pose a clear, present, and evolving threat to the public across national security, law enforcement, financial, and societal domains.

To adequately characterize the current threat landscape posed by deepfakes, we must consider the context of the varied threats, malign actors, victims, and technology. Experts across industry, academia, and government have a diverse perspective on the threat of synthetic media. Some assert that it is overblown, overstated, and technically infeasible to create a convincing deepfake, or it will simply be ineffective against a skeptical public. However, ask someone victimized by a malicious non-consensual synthetic pornography attack and they will express that the threat is very real and harmful.

The truth is that deepfake attacks are here and appear to be proliferating in domains like non-consensual pornography and in limited influence campaigns on social media platforms.

Since 2019 malign actors associated with nation-states, including Russia and China, have conducted influence operations leveraging GAN-generated images on their social media profiles. They used these synthetic personas to build credibility and believability to promote a localized or regional issue. This is not a singular incident and it seems to be a common technique now in the age of influence campaigns. Social media platforms, such as Facebook,, and other AI/ML research companies, such as Graphika, were able to detect these profiles and assess the images were AI-based and synthetically generated. It was a great success in detecting these synthetic personas, yet it wasn't always timely and it is nearly impossible to say how many other social media profiles using GAN-generated images have not been detected. Below are some specific examples of synthetic content used as part of influence campaigns:

- From 2020 to 2021, social media personas with GANs-generated images criticized Belgium's posture on 5G restrictions, in an apparent effort to support Chinese firms attempting to sell 5G infrastructure.[24]
- In 2021, FireEye reported cyber actors used GANs-generated images in social media platforms to promote Lebanese political parties.[25]
- Multiple influence campaigns conducted by cyber actors associated with nation-states used GANs-generated images targeting localized and regional issues.[26, 27, 28]

Non-consensual pornography emerged as the catalyst for proliferating deepfake content, and still represents a majority of AI-enabled synthetic content in the wild.

- In October 2020, researchers reported over 100,000 computer-generated fake nude images of women created without their consent or knowledge, according to Sensity AI, a firm that specializes in deepfake content and detection. Some of these nude images apparently depicted under-aged individuals as well. The creators used an ecosystem of bots on the messaging platform Telegram to facilitate sharing, trading, and selling services associated with deepfake content.
    - Sensity AI found that approximately 90-95% of deepfake videos since 2018 were primarily based on non-consensual pornography.[29,30,31]

- In 2021, AI Dungeon, which uses synthetically generated text, was found to generate text depicting the sexual exploitation of children. This was based on user input and relied on vast training data, however, it demonstrated the unintended consequences of AI/ML-based content generation with few constraints. AI Dungeon relied on OpenAI's GPT-3, an auto-regressive language model. This model can be implemented in a number of different capabilities to include natural language generation, text-to-image generation, translation, and other text-based tools.[32]

- In December 2019, a journalist investigating deepfake pornography and the technology behind it joined a face-swapping marketplace and paid a synthetic content creator to digitally insert her face into pornography. Instead of submitting a multitude of photos to create content, the journalist found that a 15 second Instagram story, consisting of 450 individual frames, provided all of the frames needed to create this content. The journalist sent in a 13 second video of herself talking to a front-facing camera and a link to the Pornhub video where she wanted her face inserted.[33]

Additional Factors Affect the Threat Landscape

A variety of factors affect the threat landscape, including the advancing capabilities of AI/ML-generated synthetic content; legal frameworks; norms and thresholds agreed to by nation-states; the opportunity to use synthetic content to carry out fraud schemes; and the susceptibility of the public to believe what they see. Not all malign actors will be affected by every one of these factors. Cyber criminal actors involved in financial fraud schemes will not care about norms governing the use of synthetic content by nation-states, for example. Nation-states may view a dramatic deepfake attack as an escalation, and thus while they may be among the most capable of actors, they may also be the least likely to deploy a

deepfake attack. Comparatively, cyber criminals and other individuals will likely be undeterred from creating synthetic media. The "high-impact, low probability" attacks will likely be outnumbered by the "low-impact, high probability" attacks, but by no means is being victimized by non-consensual pornography, for example, low-impact to that individual. The general public may be much more resilient against deepfake content after they've experienced an initial attack, and therefore malign actors might prefer to wait for a 'big score' before they attempt a high-impact deepfake. These actors may assess that the window of opportunity is closing as the public gradually builds resilience against synthetic media and may be incentivized to go for that 'big score' sooner than later.

Where might the deepfake threat go in the future?

Deepfakes continue to pose a threat for individuals and industries, including potential large-scale impacts to nations, governments, businesses, and society, such as social media disinformation campaigns operated at scale by well-funded nation state actors. Experts from different disciplines whose research interests intersect at deepfakes tend to agree that the technology is rapidly advancing, and the high cost of producing top-quality deepfake content is declining. As a result, we expect an emerging threat landscape wherein the attacks will become easier and more successful, and the efforts to counter and mitigate these threats will need orchestration and collaboration by governments, industry, and society.

DEEPFAKE SCENARIO EXAMPLES

The potential threat of deepfakes can be better understood within the context of specific scenarios. This section provides the reader with examples of scenarios in which deepfakes might play a role. While this list includes many examples, it is far from exhaustive.

How did we choose these scenarios?

We identified three major categories within which to assess the threat of deepfakes and synthetic media: national security & law enforcement; commerce; and society. For each of these categories we envisioned scenarios wherein we juxtaposed the future state of deepfake technology with what we know of existing and potential threat dynamics. The following section examines several such scenarios for each category, centering our analysis around four key questions – goals, use, expected gains and methodology or steps.

What might a deepfake attack look like?

National Security & Law Enforcement Scenario 1: Inciting Violence

An example scenario involving a deepfake video potentially fueling unrest and violence was suggested by Professor Danielle Citron in testimony to the House Permanent Select Committee on Intelligence in 2019. In this scenario a malign actor produces a deepfake video of the Commissioner of the Baltimore Police Department (BPD) endorsing the mistreatment of Freddie Gray, an African American who died in police custody in 2015.[34]

The scenario described by Professor Citron could unfold in the following way. The malign actor, having decided to incite violence by means of publishing a deepfake, first performs research on the Baltimore Police Department to gather still images, video, and audio of the Commissioner to use as training data. This could come from past press conferences and/or news stories.

Next, the malign actor uses the gathered information to train an AI/ML model to mimic the likeness and voice of the Commissioner. With a trained model, the malign actor creates a deepfake video of the police Commissioner endorsing the mistreatment of Mr. Gray, staged as a private discussion to lend credibility to the statements, which might also include inflammatory remarks. The malign actor could then anonymously posts the video to social media sites and draws attention to it using fake social media accounts.

National Security & Law Enforcement Scenario 2: Producing False Evidence About Climate Change

In this scenario, Chinese satellites would capture images of the Antarctica and the surrounding ice sheet. Practitioners would use AI/ML models to generate features that make it appear that ice growth has increased or stayed steady instead of decreased. China then uses this synthetically generated satellite imagery to convince the United Nations and others to delay or cancel implementation of stricter climate agreements that would be unfavorable to China's economic development. If unsuccessful, China could use this fake data to credibly dispute the need for these agreements and ignore environmental restrictions such as greenhouse gas emissions. Although US and other NGO satellites may have other data to contest China's submission to the UN, this could cause delays, confusion, and undermine global agreements.

National Security & Law Enforcement Scenario 3: Deepfake Kidnapping

In this scenario, a criminal gang operating in a tourist location in Mexico conducts targeted and opportunistic fraud schemes against victims using synthetic images and video to depict someone in a situation of captivity. The malign actors wouldn't actually kidnap someone but would use images and information they find either online or from a stolen device to conduct the fraud scheme. The malign actors would then contact the family of the target and demand ransom. They would show the "proof of life" of the victim in a hotel room, possibly bound and blindfolded. The malign actors could also send follow-up images of the victim with indications of injury to place more pressure on the victim's family to pay the ransom. The victim themselves would not be in any direct harm and may be completely unaware of this taking place.

National Security & Law Enforcement Scenario 4: Producing False Evidence in a Criminal Case

In this scenario, a wealthy criminal defendant, who is accused of murder in a building he owns based on a number of pieces of evidence including latent fingerprints, hair DNA and motive, has objected to the use of identity verification from videos captured in the building lobby based on low resolution and lack of clarity on the face image. As an alibi, he is submitting to the court video imagery from another location in the building that irrefutably puts him elsewhere at the time the crime took place.

The goal in this case is to weaken the biometric evidence and offering contradictory proof that the defendant has a clear alibi. The deepfake content here is the submitted video itself. It is not only the timestamp that can be modified to provide an alibi, but unique circumstances can be created in the video so that it looks genuine. The deepfake video, indirectly, also could also render otherwise strong evidence (such as biometrics) circumstantial, due to the specifics of the building and the scene.

Commerce Scenario 1: Corporate Sabotage

In this scenario we consider the use of deepfake technology to spread misinformation about a company's product, place in the market, executives, overall brand, etc. This approach is designed to negatively affect a company's place in the market, manipulate the market, unfairly diminish competition, negatively affect a competitor's stock price, or target prospective mergers & acquisition (M&A) of a company.

Commerce Scenario 2: Corporate Enhanced Social Engineering Attacks

In this scenario we consider the use of deepfake technology to more convincingly execute social engineering attacks.

First, a malign actor would conduct research on the company's line of business, executives, and employees. He identifies the Chief Executive Officer (CEO) and the Finance Director of the company. The malign actor researches a new joint venture that company announced recently. He utilizes Ted Talks and online videos of the CEO to train the model to create a deepfake audio of the CEO. The malign actor conducts research on the Finance Director's social media profiles. He sees that he posted a picture of a baby and a message it's hard to return to work. Next, the individual would place a call to the Finance Director with a goal to fraudulently obtain funds. He would ask the Finance Director about how he is doing returning to work and about the baby. The Finance Director answers his phone and recognizes his boss's voice. The malign actor directs him to wire $250K to an account for the joint venture. The funds would be wired and then the malign actor would transfer the funds to several different accounts.

REAL-TIME VOICE ALTERATION
(FOR ON-PHONE SCAMMING)

Commerce Scenario 3: Financial Institution Social Engineering Attack

In this scenario, the malign actor decides to employ a deepfake audio to attack a financial institution for financial gain. Next, she conducts research on the dark web and obtains names, addresses, social security numbers, and bank account numbers of several individuals. The malign actor identifies the individuals' TikTok and Instagram social media profiles. She utilizes the videos posted on social media platforms to train the model and creates deepfake audio of targets. The malign actor researches the financial institution for the verification policy and determines there's a voice authentication system. Next, she calls the financial institution and passes voice authentication. She is routed to a Representative and then utilizes the customer proprietary information obtained via the dark web. The malign actor tells the Representative that she was unable to access on her account online and needs to reset her password. She was provided a temporary password to access the online account. The malign actor gains access to her target's financial accounts. The malign actor wires funds from the target's account to overseas accounts.

Commerce Scenario 4: Corporate Liability Concerns

If deepfakes become convincing enough and ubiquitous enough, companies may be at increased legal risk due to affected consumers' seeking damages and compensation for

financial loss due to ensuing breaches, identity theft, etc. Additionally, consumers or patrons could fabricate an incident (e.g. a slip and fall in a grocery store) to defraud a company.

The first scenario would not be an attack, per se, but more of a consequence put into motion by previous deepfake attacks.

In the second scenario, we consider a deepfake video designed to make a company seem liable for a product that malfunctioned and caused an injury to defraud the company. The malign actor would conduct research and identify a product with a history of causing a physical injury. Specific details are extracted from previous incidents to include in the deepfake video. Next, the state tort laws would be reviewed to determine the likelihood of a quick settlement. The malign actor finds videos on YouTube and other social media platforms of the same product malfunctioning and injuring people. The face swap model is trained and the deepfake video is created. The malign actor releasing the video on social media and receives an outpouring of support. The company's social media team sees the post and reaches out to the malign actor. The pressure form social media is mounting for the company to take accountability. The request for compensation for the injuries is made to the company. Will the company have the ability to determine if the video is a deepfake before they pay the malign actor? Does the company have the time to investigate the video? Is it worth paying the malign actor and reducing the damage to the company's brand?

Commerce Scenario 5: Stock Manipulation

In this scenario, we consider a deepfake generated to manipulate the stock market and allow the malign actor to make an illicit profit.

A malign actor wishes to make a quick profit through stock manipulation. The actor thoroughly researches the stock and purchases it at a low price. He creates several deepfake profiles on stock market forums such as Reddit and Stockaholics. The profiles show that the users are employees of the company. Posing as these employees, the actor posts comments about a pending "major" announcement. Having identified the company CEO, the actor trains a model of the CEO's speech based of interviews which aired on various television and radio programs. The actor creates an audio deepfake of the CEO discussing the pending "major" announcement and posts it on social media, along with a link to the audio on stock market forums. The malign actor monitors forums and sees a huge spike of activity confirming his deepfake audio is working. The stock increases in price by 1000 percent and the malign actor cashes out before the stock drops. This could cause other investors to lose money and impact the company's reputation. The company may make a statement that the audio of the CEO was fake. The investors may look to the company to make them whole for any losses suffered.

Society Scenario 1: Cyberbullying

In this scenario we consider a deepfake generated to depict a target in a situation which would damage their reputation or impact their access to groups, services, or benefits, perhaps by depicting the individual engaged in criminal behavior.

The attacker wishes to undermine the reputation of the target, which may have the secondary effect of enhancing the status of another preferred by the attacker. In a well-publicized recent incident in Pennsylvania, a woman attempted to damage the reputation of

cohorts of her daughter who were in competition for limited spots on a cheerleading squad.[35]

In this scenario a deepfake video depicting the target engaged in criminal behavior is produced and sent to individuals in positions of authority over the target's activities. Based on the video, these authorities restrict or remove the target from participating in certain activities.

Cyberbullying Scenario – Deepfakes Implicated

Cyberbullying is a common issue especially with younger generations due to the high usage of social media. Rumors can be easily spread through social media and online platforms, which, when coupled with fake images or videos to suggest the rumor is true, can make the rumor more believable. This can ruin reputations and cause psychological effects that may lead to victims hurting themselves.

In March 2021, international news brought to light an incident once charges were filed that involved using alleged deepfakes as a cyberbullying tactic. In Pennsylvania, a mother allegedly manipulated images and videos of her daughter's cheer squad teammates. These alleged deepfakes showed members of the cheer squad drinking, vaping, and posing nude, all actions that could get them cut from the cheerleading squad. Several of the victims came forward about the cyberbullying and one victim claimed the mother went as far as encouraging suicide, furthering the harassment outside of the alleged deepfakes.

However, by May 2021, the deepfake accusation had been abandoned as it could not be proven that the video evidence was falsified. Synthetic-media researchers noted that the videos did not carry traditional manipulated signatures such as artifacts around the face to suggest it had been altered, but it did have realistic details that would be difficult to fake such as the vapor cloud one individual exhaled. Multiple digital forensics experts stated the videos appeared to be authentic and it was unlikely they were actually deepfakes.

Another issue is how easily accessible tools and resources are to the public to create deepfakes. There are mobile applications and web applications that are either free or low cost to download and create deepfakes which could be then used maliciously in scenarios like cyberbullying. The use of deepfakes in cyberbullying cases will likely increase and become more of a threat as time goes on, especially for younger generations who frequently use technology and social media.

Society Scenario 2: Deepfake Pornography

In this scenario we consider non-consensual deepfake pornography. Karen Hao, MIT Technology Review Senior Editor, stated "the biggest threat is to women and vulnerable populations[36]. By far 95% of deepfakes are of nonconsensual porn of women. Individual level is the highest threat[37]". This number includes "anyone whose image has been

captured digitally" and posted on the internet. This applies to virtually every woman in the country – if not the world – and, therefore, poses an exponentially larger risk[38]

At present, anyone with a social media profile is fair game to be faked[39]. Consider a scorned boyfriend who wishes to blackmail a girlfriend who has expressed the desire to break up so she can date someone else. The boyfriend (attacker) wants to frighten the girlfriend (victim) into staying with him by threatening to publish naked photos of the victim if she follows through on the break up.

When the victim refuses to stay in the relationship, the attacker collects headshots of the victim from photos he took when they were a couple, as well as from one of her social media accounts. He then collected several pictures of naked women from the Internet.

The attacker then uses a free app called "Reflect" to graft the victim's head on to the body of one of the naked women whose pictures he collected online. It might take the attacker as little as five minutes to create this picture, even if he is a novice.

Having created the picture and cropping out any contextual information which might indicate it is a fake (e.g., a watermark from the app, or information indicating a location the victim has never visited), the attacker sends the picture to the victim's parents and friends. He also posts the photo to social media to further embarrass the victim. Although the victim knows the photo is a fake, she suffers humiliation because some people will believe the photo is authentic.

CASE STUDY BREAKOUT – NOELLE MARTIN

Our team interviewed Noelle Martin, Australian Activist and Lawyer, who has been severely impacted by nonconsensual deepfake technology. In 2016, when she was 17 years old, Ms. Martin discovered that a selfie of her face had been superimposed into a pornographic image and distributed to several porn sites. Ms. Martin stated she was targeted a second time because she spoke up and a deepfake nonconsensual pornographic video was made of her. Ms. Martin stated she believed the video was used as a weapon to try silence her.

Ms. Martin also stated that the deepfake nonconsensual pornographic video was emailed to her and posted online on various sites. "…it was of a woman on top of a man, having sex with him, her whole naked body was visible, her eyes looked straight into the camera as her body moved and her face reacted to the activity. Except it wasn't the face of a stranger… it was my face[40]." Ms. Martin notes that deepfake pornography has lifelong effects on an individual's reputation, dignity, employability, and interpersonal relationships. During the interview, she stated that she was unable to obtain employment in her field of work. Ms. Martin believes the difficulty she has faced in her job search was attributed to the deepfake pornographic images and videos of her.

Ms. Martin reflected on the never-ending struggle she faced when trying to identify the individuals responsible for the deepfake attacks against her. Five years later, Ms. Martin stated she still did not know the malign actor(s) that targeted her. Videos still pop back up on sites and she has found it very hard to remove videos permanently. Ms. Martin noted

that there were no laws that addressed deepfake pornography. Although many existing legal claims might provide adequate redress in highly specific circumstances, none are sufficient to address deepfake pornography at large, revealing the need for a new solutions[41]. Ms. Martin stated that all stakeholders need to come to the table to address this threat, that there should be global collaboration between law enforcement agencies, and that victims need more mental health support.

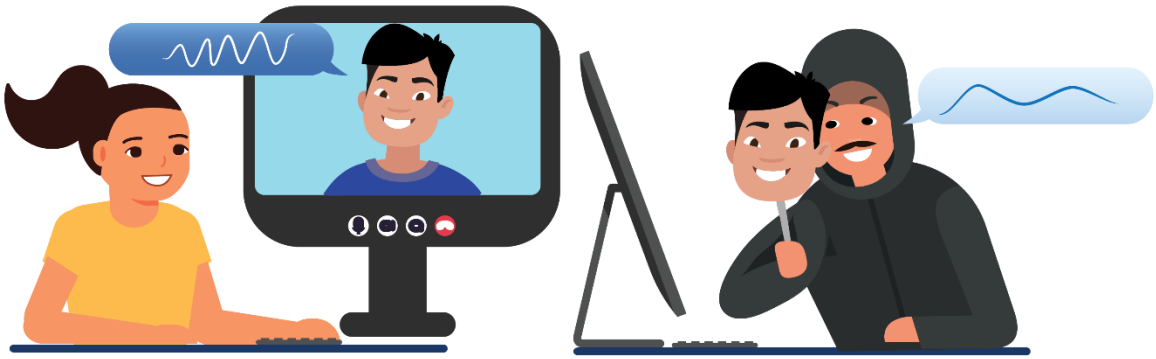
Society Scenario 3: Election Influence

In this scenario we consider a deepfake used to spread disinformation around the time of an election. In the run-up to the election, a group of tech-savvy supporters of Candidate A could launch a disinformation campaign against Candidate B. In the scenario, malign actors may take advantage of audio, video, and text deepfakes to achieve their objectives. While individual audio and video deepfakes are likely to create more sensational headlines and grab people's attention, the threat of text deepfakes lies in their ability to permeate the information environment without necessarily raising alarms.

Another key use of text deepfakes is controlling the narrative on social media platforms. This approach could heighten societal tensions, damage the reputation of an opponent, incite a political base, or undermine trust in the election process.

Society Scenario 4: Child Predator Threat Scenario

An additional threat scenario enabled by deepfakes and synthetic media involves a child predator who uses the technology to create an avatar that appears to be much younger in order to target children in online conversations. The predator would not necessarily need to recreate a specific, targeted individual, but would just need to create a child's face and voice that are realistic. If encountered online in a chat room or other social media environment, the potential victim might not be able to detect the deepfake.



REAL-TIME VOICE AND VIDEO ALTERATION (FOR ONLINE CHAT)

Scenarios – Summary

We are already seeing that the unbounded use of tools for audiovisual manipulation often has a negative impact on women, people of color, and those questioning powerful systems[42]. The fact that deepfake technology will be accessible on a large scale to many people is a challenge. Each technical approach described was previously only available to experts, but in the context of technological advancement and widespread social media use, these approaches are more accessible to amateurs and their outputs reach larger scales at higher speeds[43]. Regardless of the intention of those who create videos, the effects wrought by technology matters. AI-generated pornographic videos highlight a set of interconnected problems regarding the "democratization" of a technology[44].

CONSIDERATIONS FOR MITIGATION

Commonalities Across Scenarios

Despite the wide range of possibilities for an attack using deepfakes, the scenarios we present in this paper follow a generalizable progression that could be used to engage with the appropriate stakeholders and help inform mitigation measures. These steps are similar to those outlined in a paper by Jon Bateman[45] and include:

1. **Intent:** Any attack which utilizes a deepfake must begin with a malign actor choosing to make an attack against a given target.

2. **Researching the Target:** In this step, the malign actor performs research on the target to gather still images, video, and/or audio. This could come from a variety of sources including search engines, social media sites, video sharing sites, podcasts, news media sites, etc.

3. **Creating the Deepfake, Part 1 - Training the Model:** The malign actor uses the gathered information to train an AI/ML model to mimic the likeness and/or voice of the target. Depending on the resources and technical sophistication of the actor, this could be done using custom models or commercially available applications.

4. **Creating the Deepfake, Part 2 – Creating the Media:** The malign actor creates a deepfake of the target doing or saying something that they never did. This could be done using the actor's hardware, commercial cloud infrastructure, or a third-party application.

5. **Disseminating the Deepfake:** The malign actor releases the deepfake. This could be done in a targeted fashion to an individual or more generally to a wide audience through methods such as sending through email or posting to a social media site.

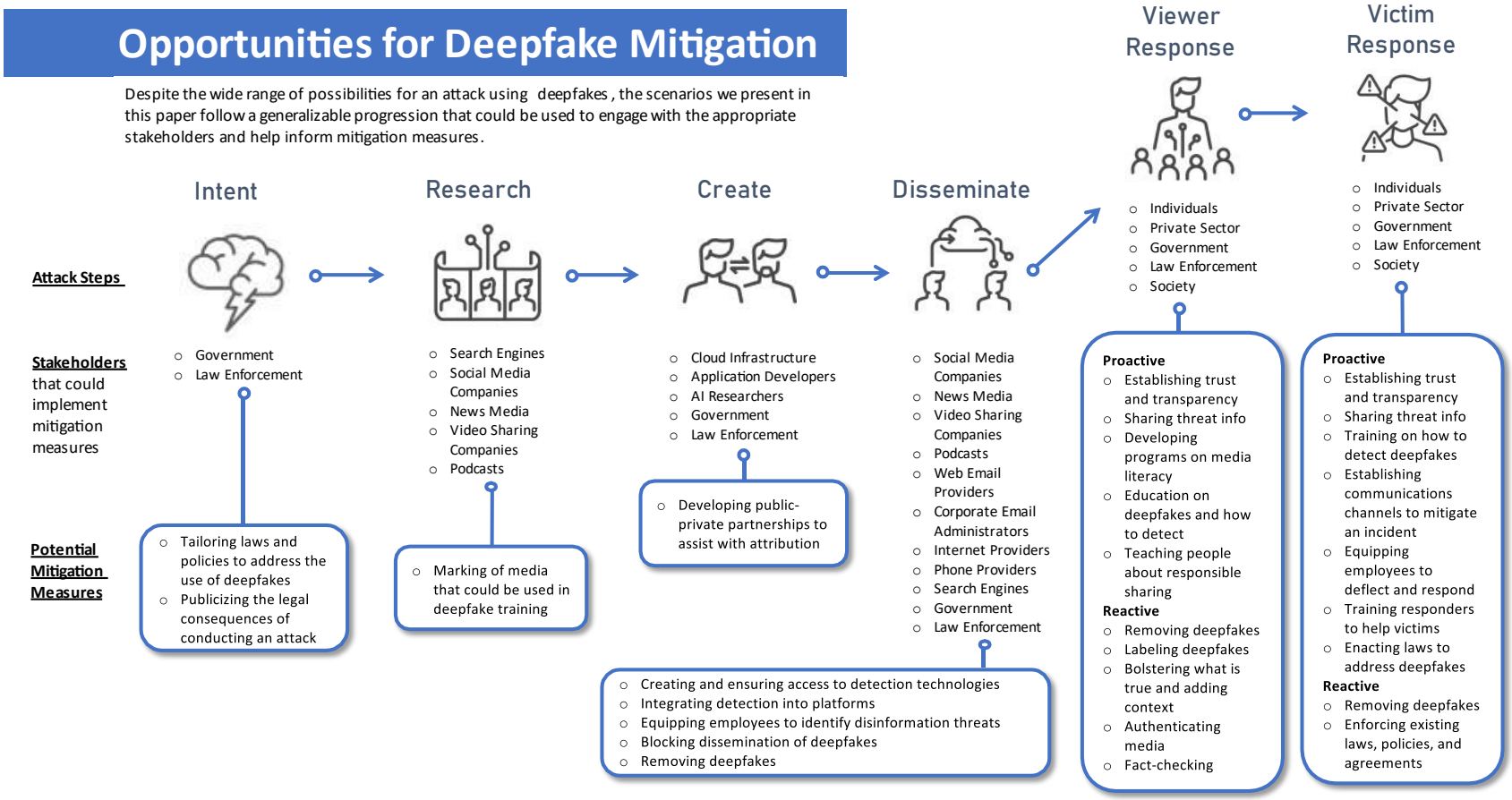
6. **Viewer(s) Respond:** Viewers react and respond to the contents of the deepfake.

7. **Victim(s) Respond:** The victim of the deepfake reacts and responds, often in the form of "damage control." Although the victim is also a "viewer," their response could be very different due to their unique role in the attack.

Opportunities for mitigation are available at all these steps.

Mitigation Opportunities

Due to the complexity and unpredictability of the issue, mitigation measures for deepfakes must be broad-based, utilizing the widest possible range of available human-centered and technological solutions.



At present, and in what we might call the earlier stages of the threat of deepfakes, mitigation tactics tend to focus on the development of technological solutions, primarily automated deepfake detection. However, as deepfakes progress and become more pervasive and ubiquitous, this single-minded approach will no longer be adequate, placing individuals and organizations on the defense and in a constant battle to catch up with the latest threat. This reactive approach would be both inefficient and needlessly risky.[46]

In pursuing a proactive approach, a full complement of mitigation measures might address individual steps in the progression of a deepfake and should ideally be inclusive of the following elements:[47,48 ,49 ,50]

Intent Phase Mitigations

- Laws, policies and regulations that make it a crime to disseminate malicious content;
- Publicizing the criminal and civil punishments that may apply if malicious content is shared could lead to some reduction in the chance of malign use;
- Social pressure on individuals identified as having produced malign content may also dissuade some from creating malign content. (Community self-policing?)
- Policy and law: It is clear that appropriate policy and laws tailored to address the emerging threat of deepfakes must be researched, developed and implemented.
- Refer to the text box on "Civil Litigation" for more.

Civil Litigation

Recently, several bills addressing deepfakes have been introduced in the US Congress and across state legislatures. The state of Virginia extended its revenge porn law to include criminalizing nonconsensual deepfake pornography; Texas enacted laws criminalizing deepfakes that interfere with elections; California passed two laws that allows victims of nonconsensual deepfake pornography sue for damages and public office candidates the ability to sue individuals or organizations that create or share election-related deepfakes within 60 days of an election; and Maryland, New York, and Massachusetts are considering their own specific approaches to legislating deepfakes.[5152] Nonetheless, it is argued that state laws are not the best way to solve the problem of deepfakes since each legislation would target different aspects of deepfakes and only apply to specific states.[53] However, the need to rush to enact deepfake-specific laws might not be necessary.

Law enforcement officials can rely on existing criminal laws until deepfake-specific federal laws are passed. In some cases, victims can pursue a civil suit by claiming extortion, harassment, copyright infringement (i.e. using images protected by copyright law without permission), intentional infliction of emotional distress, defamation, false advertisement, and false light (i.e. a form of invasion of privacy). For example, if a deepfake is used in an extortion scam, extortion laws could apply, and in an event where deepfakes were used to harass someone, harassment laws could apply. Also, if a malign actor publishes offensive false information about a person and implies that it is true (e.g. photo manipulation and embellishment), a victim can purse a false light invasion of privacy civil claim.[54]

Research (and other Pre-Dissemination) Phase Mitigations

- Organizational planning: Establishing and nurturing effective communications structures and channels to mitigate deepfakes incidents, when they inevitably occur, will be key. Just as most organizations have PR plans in place in case of a reputational crisis way ahead of said potential crisis, so should they have plans in place to monitor and control their own narrative and avert potential disaster caused by misinformation, in its various guises. Specific tactics in such a plan could include developing a disinformation response policy as part of an overall information security policy and having dedicated monitoring and reporting of information, as well as disinformation, on social media and other outlets.
    - Organizations and individuals which may be targets – political and commercial – can also be proactive in monitoring and curating their multimedia output, especially any broadcast or public content. The reasons for close curation is twofold: (1) When an existing piece of content is repurposed in a deepfake, it may be possible to quickly identify the original source content and offer that as the "authentic" media; and (2) Some of the best deepfake detection models[55] today use a process whereby a speaker's facial movements in a video can be measured to determine if they are consistent with prior instances of authentic speech. When a mouth replacement or puppet technique is used to create a deepfake, the overall facial movements in the altered video will differ enough from the true face that it can be detected as a deepfake.
- Training and awareness: Using the example of phishing prevention and mitigation, employers should consider investing resources in equipping employees with the knowledge and skills to serve as on-the-ground "first responders" to deflect and/or report on disinformation and related threats in the workspace, including deepfakes. Other training measures specific to law enforcement and other officials, could focus on training them to help victims mitigate the impact of deepfakes attacks on and their reputation, health, and welfare. See below for more.

Creation Phase Mitigations

- Organizations and individuals who develop models utilized in deepfake creation should also consider their responsibilities when it comes to mitigation. Individuals and organizations who are concerned that their developments not be used in an irresponsible or criminal manner could take steps that would make it easier to detect when their model was used. For example, the developers may be aware of a weakness or signature of their model which makes it easy to detect. Rather than hide that fact, the developers could release the signature with their code as a sign of their interest in being a responsible member of a technologically advanced society.
- Creators of deepfake content could also mark their creation as a "deepfake."

Dissemination Phase Mitigations

- Partnership development: Partnerships among industry, academia, and law enforcement, among other entities, would hopefully speed up the process of detecting, labeling, and removing non-consensual images and other defamatory synthetic media, when they occur.
- Detection and other technological innovation: Since the technology may be used for entertainment, education, and protected speech, purposes, deepfake detection alone cannot constitute an entire mitigation protocol. Nevertheless, the overall importance and impact of such technological tools cannot be discounted. Successful detection allows for early intervention and mitigation. Social media platforms, Internet service providers, and other communication systems providers – those who provide the pipes through multimedia flows – are best positioned to identify the nature of content that they are transmitting. Technologies such as Microsoft's PhotoDNA – which identifies copies of previously identified digital images - have been deployed by some of these providers to stem the flow of child sexual abuse materials (CSAM), so there is precedent.
- As detection tools are developed, they can be shared with social media companies, Internet Service Providers, and communication system providers, and made available as open source tools.
    - The creation and distribution of deepfake content which can be used to train both humans and models in detection could also assist.
- On the flip side of detecting synthetic content, there is also the opportunity to promote authentication measures. More specifically, individuals and organizations can take steps to demonstrate and verify the authenticity of the media they create and consume.
    - For example, in 2019, the Content Authenticity Initiative (CAI) was created[56]. The CAI describes itself as "…a community of media and tech companies, NGOs, academics and others working to promote adoption of an open industry standard for content authenticity and provenance." These standards would allow users to demonstrate the provenance and attribution of their media content in a manner accessible to all who use the standards. In this way, consumers could check media for a "seal of authenticity," allowing for greater trust in the content.
    - While the CAI would offer open standards available for adoption by any user, commercial service provider can also offer similar capabilities. Truepic is one such company.
- Finally, success on this front can also be achieved through secure communication channels where users control all of the content (i.e., closed networks).

What can we do to increase trust in real-time interactions or media?

Increasing the public's trust in real-time interactions and media is a long-term prospect, but nonetheless a critical step to protect society and institutions from disinformation. A renewed adherence to security protocols such as 2-Factor authentication and device-based authentication constitute an elemental first step in this process. Additionally, it would be advantageous to pursue a strategy of investment in strengthening our democratic and media institutions and newer and emerging technology. Block chain authentication is one such stand-out possibility, as it holds great potential to standardize and promote verification and authenticity, thereby creating a trusted space and inspiring consumers' confidence in what is seen and heard.

Viewer Phase Mitigations

- Societal education: Polices and programs need to be set up to offer greater educational outreach to the public, addressing the issue of misinformation resistance and strengthening the public's ability to discern fact from fiction. Additionally, it would be advantageous to develop and prioritize a program of early education in media literacy and critical thinking skills.

How can we determine what is real and what has been manipulated? What can we do to improve detection? Can we educate the public to detect a deepfake? Should we?

As technology advances, it will become increasingly difficult to identify manipulated media. There are commercial tools that can be used to help detect fake media, but these tools will need to be constantly retrained and updated to detect variables and several manipulation aspects. Every tool may vary in what they quantify as a deepfake as well, which will affect the type of media manipulation it would flag. Depending on how sophisticated the deepfake is, the public may be able to detect it with their own eye or forensic experts can analyze the content more closely. It would be more efficient to have AI/ML tools do that work up front instead of humans doing it manually but improving deepfake detection will be an all hands-on deck situation. As with any AI/ML model, it will have to be trial and error to see what the tool can recognize and continually running fake media through it to see what it misses. Then the model can be adapted to lessen the possibility something will slip through the cracks. From the society aspect then, individuals should be checking credibility of media before passing it on so the spread of misinformation does not grow.

Experts in the field of AI/ML, including the Partnership on AI (PAI), have suggested that to improve detection, a sort of paradigm shift should occur from focusing on detecting what is

fake to bolstering what is true about the media and adding context to the media. This will empower individuals to explore the authenticity of media by using context clues or metadata about where the media originated to help determine if it is real. PAI has conducted interviews that suggest individuals do not want to be told what to believe or taught in a condescending way of what is real or not, they want to figure it out for themselves. This could also improve trust in institutions if individuals can legitimize the media that is being are shared.

To empower individuals to authentic media, society will have to be educated on deepfakes. Individuals may not know the true meaning of what deepfakes are or the extent of harm they can cause, but if they can be taught what to look for, they may be able to detect it on their own. Educating society on deepfakes and how easily they can be created is important because due to society's social media usage, fake media can be shared very easily without a second thought to if it is real or not. The issue with education is that society will have to want to learn about deepfakes. There are going to be individuals who may not have an interest in learning about them and just believe whatever narrative they want to believe. There is a chance though that if society cannot be taught to analyze and learn about deepfakes, they can learn to just not share everything they see.

An individual should look for the following signs when trying to determine if an image or video is fake:

- Blurring evident in the face but not elsewhere in the image or video (or vice-versa)
- A change of skin tone near the edge of the face
- Double chins, double eyebrows, or double edges to the face
- Whether the face gets blurry when it is partially obscured by a hand or another object
- Lower-quality sections throughout the same video
- Box-like shapes and cropped effects around the mouth, eyes, and neck
- Blinking (or lack thereof), movements that are not natural
- Changes in the background and/or lighting
- Contextual clues – Is the background scene consistent with the foreground and subject?

An individual should look for the following signs when trying to determine if an audio is fake:

- Choppy sentences
- Varying tone inflection in speech
- Phrasing – would the speaker say it that way?
- Context of message – Is it relevant to a recent discussion or can they answer related questions?
- Contextual clues – Are background sounds consistent with the speaker's presumed location?

An individual should look for the following signs when trying to determine if text is fake:

- MIsspellings
- Lack of flow in sentences
- Is the sender from a known number or email address?
- Phrasing – would the legitimate sender speak that way?
- Context of message – Is it relevant to a recent discussion?

What can we do if it has been determined that media is "fake" or manipulated? Prevent content from being posted or delivered? Remove content? Label content?

According to Kathryn Harrison, the founder and CEO of DeepTrust Alliance, there is little that can currently be done after a deepfake is discovered. One option would be to determine how much the media has been manipulated, but this could be time consuming and expensive. Harrison suggests hiring a lawyer to help look through Terms of Services on platforms in which the deepfake was shared. This will also be time consuming as each platform will have different terms in place and lawyers would have to negotiate IP and copyright protections to have content taken down from the site. Deepfakes could in some instances be raised up to the law enforcement level for reporting, but because the threat is so new, there are not many regulations in place. Most law enforcement agencies will not have protocols or tools in place to handle deepfake cases and sometimes it may be out of their jurisdiction.

Prominent AI/ML companies and social media platforms could be successful avenues to focus on deepfake mitigation. AI/ML tools could update models over time to constantly adapt with the evolving technologies to identify deepfakes. However, trying to get society as a whole to take the time to run media through these tools to determine media authenticity will be an issue. Social media platforms could integrate these tools to flag content and minimize the spread of deepfake content. If the major platforms work together with AI/ML companies then deepfake content could be identified from the get-go and never be posted or at least removed quicker from their sites before it goes viral to the public.

There has been ongoing research and support dedicated to deepfake detection and other proactive mitigation efforts to alleviate the threat posed by deepfakes. However, more emphasis should be placed on supporting victims (i.e. individuals or businesses) who are recovering from deepfake attacks. According to Noelle Martin, a law graduate and activist in Australia, she was 17 years old when she discovered that a picture of her had been photoshopped onto pornographic images and distributed across porn sites. The photoshopped images later evolved into deepfake videos, which were emailed to her and posted to online websites. Noelle also stated that being a victim of deepfakes can have

lifelong effects on an individual's reputation, dignity, employability, and interpersonal relationships. At one point, due to the nature of the deepfake videos made of her, she found it difficult to find employment in the legal field. Noelle chose to speak out against the attack; however, speaking out led to more online attacks.[57] *Still to this day Noelle does not know who targeted her or why she was targeted, and at the time of the attack there was no legislation to deal with the issue. So what can be done to help victims recover from deepfake attacks?*

Victim Phase Mitigations

- Victims of synthetic media attacks, especially non-consensual sexually explicit media attacks, often note the difficulty of removing content from all potential sources. Victims describe the recurring nightmare of having the same content appear on multiple sites and the frustration of having to solicit formal government action before it can be removed. Noelle Martin proposed a scenario in which responsible government agencies take a proactive role in seeking out content which has been formally judged as malign in nature and worthy of removal. Governments have worked with Internet service providers and social media companies to identify hash sets of CSAM material. It is technologically feasible to repeat this process for malign deepfake content.

Reporting Deepfakes

There are ways victims can report deepfake attacks. Victims of deepfakes could:

- *contact law enforcement officials who could possibly help victims by conducting forensic investigations using police reports and evidence gathered from victims;*
- *contact the Federal Bureau of Investigations (FBI) and report incidents to local FBI offices or the FBI's 24/7 Cyber Watch at CyWatch@fbi.gov;*[58]
- *utilize the Securities and Exchange Commission's services to investigate financial crimes*
- *report inappropriate content and abuse on social media platforms (i.e. Facebook, Twitter, Instagram, etc.) using the platforms' reporting procedures; and*
- *if a victim is under 18 years of age, incidents can be reported to the National Center for Missing and Exploited Children via their cyber tip line at https://report.cybertip.org.*

Resources Available for Victims

According to Bobby Chesney and Danielle Citron, victims may find it challenging taking the civil liability route if convincing evidence is unavailable, and even if a malign actor is identified, it may be impossible to use civil remedies if the malign actor is outside of the United States or in a jurisdiction where local legal action is unsuccessful.[59] *However, there are several resources available for victims of online abuse that could possibly support them*

in other ways. Organizations who are dedicated to helping victims include but are not limited to:

- *Cyber Civil Rights Initiative, an organization that provides a 24-hour crisis helpline, attorney referrals, and guides for removing images from social media platforms and other websites;*[60]
- *EndTab, an organization that provides victims, including universities, law enforcement, nonprofits, judicial systems, healthcare networks, with resources for education and reporting abuse;*[61]
- *The National Suicide Prevention Lifeline, a national network of local crisis centers that provides free and confidential emotional support for people in distress;*[62]
- *Cybersmile, a non-profit anti-bullying organization that provides expert support for victims of cyberbullying and online hate campaigns;*[63]
- *Identitytheft.gov, the federal government's one-stop resource for identity theft victims;*[64]
- *Withoutmyconsent.org, a non-profit organization that provides guides for preserving evidence that could be used in a civil suit;*[65]
- *Google's Help Center, a resource available via Google that enables victims to remove fake pornography from Google searches;*[66] *and*
- *Imatag, a company who offers tracking and image/video monitoring services.*[67]

FINAL THOUGHTS

The Liar's Dividend and the Spectre of Weaponized Distrust

This paper has considered the possible consequences resulting from the malign use of a deepfake, but another threat exists that may be counter-intuitive to many readers. In this concluding section, we consider the possible consequences not from the use of a deepfake, but from the mere possibility of it being used. The mere existence of deepfakes could undermine the primacy of credibility and authority of traditional social institutions, like the press, government, and academia.

First introduced by academics Danielle K Citron and Robert Chesney in *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, we explore how 'The Liar's Dividend' could manifest from an earnest attempt to counter the threats of deepfakes.

The public may well emerge to be able to endure through the threat of malign actors creating indiscernible deepfakes that attempt to cause widespread panic or harm. Media reporting on the phenomenon of deepfakes continues to highlight this issue, and in so doing, may imbue awareness and resilience to the public. If, however, the pendulum swings too far the other way, in which the public views things not only with scrutiny, but with a default posture of doubt and disbelief, then this too could be exploited by malign actors hoping to muddle reality. This persistent threat could evoke a sense of skepticism that undermines trustworthy institutions and questions the legitimacy and authenticity of true

content and media. Malign actors could intentionally sow distrust and doubt on legitimate media, suggesting that authentic content is actually an elaborate deepfake.

In a climate where the political spectrum is polarized and adversarial, with 24 hour media cycles, a politician facing a critical scandal, for example, could simply proclaim “that event never happened. It is clearly a deepfake created by my political enemies,” evade timely accountability and preserve their reputation unbesmirched.

Sophisticated actors could synthetically reproduce an already recorded event using AI/ML technologies, insert some sort of detectable signature, in an effort to trigger a response from authentication and detection tools, calling into question whether the real content is legitimate in the first place. This could enable malign actors to point at their reproduced media and claim the event never occurred. Recording both the good and the bad of history is an important tool to evolve the decency of society. There are those today that claim the Holocaust never existed. Deepfakes could be a nefarious tool to undermine the credibility of history.

What next?

Deepfakes, synthetic media, and disinformation in general pose challenges to our society. They can impact individuals and institutions from small businesses to nation states. All may be impacted by them. As discussed above, there are some approaches which may help mitigate these challenges, and there are undoubtedly other approaches we have yet to identify. Regardless of the approach, however, for any one to be successful will require collaboration across all affected parties. It is time for there to be a coordinated, collaborative approach. Our team hopes to participate in that collaboration in the years ahead.

ACKNOWLEDGEMENTS

The AEP "Increasing Threats from Deepfake Identities" Team gratefully acknowledges the following individuals for providing their time and expertise in the course of our research:

Jon Bateman, Fellow, Carnegie Cyber Policy Initiative of the Technology and International Affairs

Paul Benda, Senior Vice President Operational Risk and Cybersecurity, American Bankers Association

Bobby Chesney, Associate Dean for Academic Affairs, University of Texas School of Law

Kathleen Darroch, Senior Vice President and Security Business Partner Manager, PNC Bank

Daniel Elliot, Information Security Architect, Network Security at Johnson Controls

Jordan Fuhr, Senior Vice President, Information Security Government and Public Policy

Candice G., Applied Research Mathematician, U.S. Department of Defense

Sam Gregory, Program Director, WITNESS

Karen Hao, Senior AI Editor, MIT Technology Review

Kathryn Harrison, Founder and CEO, DeepTrust Alliance & MAGPIE

Tia Hutchinson, Policy Analyst, U.S. Department of the Treasury Office of Cybersecurity and Critical Infrastructure

Tim Hwang, Research Fellow, Georgetown's Center for Security and Emerging Technology

Ashish Jaiman, Director of Product Management, Bing Multimedia, Microsoft

Dr. Neil Johnson, Cyber & Forensic Scientist, Pacific Northwest National Laboratory

Claire Leibowicz, Head of AI and Media Integrity, Partnership on AI

Dr. Baoxin Li, Professor, Chair of Computer Science and Engineering Program, Arizona State University

Dr. Siwei Lyu, SUNY Empire Innovation Professor, Department of Computer Science and Engineering at University at Buffalo, SUNY

Dr. Sebastien Marcel, Senior Researcher and Head of Biometrics Security and Privacy, Idiap Research Institute

Noelle Martin, Activist and Survivor of Deepfake Attack

Mike Price, Chief Technology Officer, ZeroFox

Kelley Sayler, Analyst, Congressional Research Service

Thao T., Visual Information Specialist, Federal Bureau of Investigation

Dr. Matt Turek, Program Manager, Defense Advanced Research Projects Agency

Dr. Matthew Wright, Professor of Computing Security and Director of Research for the Global Cybersecurity Institute at the Rochester Institute of Technology

END NOTES

[1] (U) | Samantha Cole | Vice | https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn/| 11 Dec. 2017 | AI-Assisted Fake Porn is Here and We're All Fucked.
[2] (U) | Samantha Cole | Vice | https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn/| 11 Dec. 2017 | AI-Assisted Fake Porn is Here and We're All Fucked.
[3] (U) | Juergen Schmidhuber |Neural Networks | DOI 10.1016/j.neunet.2014.09.003 |Jan. 2015 | Deep Learning in Neural Networks: An Overview | p 85–117.
[4] (U) | Dave Gershgorn |OneZero | https://onezero.medium.com/deepfake-music-is-so-good-it-might-be-illegal-c11f9618d1f9 | 01 May 2020 |Deepfake Music Is So Good It Might Be Illegal.
[5] (U) | Nick Statt | The Verge| https://www.theverge.com/2020/4/28/21240488/jay-z-deepfakes-roc-nation-youtube-removed-ai-copyright-impersonation/ | 28 Apr. 2020 | JayZ tries to use copyright strikes to remove deepfaked audio of himself from YouTube.
[6] (U) | Renee Diresta | Wired | https://www.wired.com/story/ai-generated-text-is-the-scariest-deepfake-of-all/ | 31 Jul. 2020 | AI-Generated Text Is the Scariest Deepfake of All.
[7] (U) | Steven Rosenbaum | MediaPost| https://www.mediapost.com/publications/article/341074/what-is-synthetic-media.html | 23 Sep. 2019 | What Is Synthetic Media?.
[8] (U) | Sudharshan Chandra Babu |Paperspace Blog| https://blog.paperspace.com/2020-guide-to-synthetic-media |2019 | A 2020 Guide to Synthetic Media.
[9] (U) | Candice Gerstner, Emily Philips, Larry Lin | NSA| The Next Wave | ISSN 2640-1797| 2021 | Deepfakes: Is a Picture Worth a Thousand Lies? | p 41- 52.
[10] (U) | Candice Gerstner, Emily Philips, Larry Lin | NSA| The Next Wave | ISSN 2640-1797| 2021 | Deepfakes: Is a Picture Worth a Thousand Lies? | p 41- 52.
[11] (U) | Britt Paris, Joan Donovan | Data & Society| https://datasociety.net/library/deepfakes-and-cheap-fakes/ | 18 Sep. 2019 | DEEPFAKES AND CHEAP FAKES.
[12] (U) | Alan Zucconi | https://www.alanzucconi.com/2018/03/14/understanding-the-technology-behind-deepfakes/ | 14 Mar. 2018 | Understanding the Technology Behind DeepFakes.
[13] (U) | Claudia Willen |Insider| https://www.insider.com/kristen-bell-face-pornographic-deepfake-video-response-2020-6| 11 Jun. 2020 | Kristen bell says she was 'shocked' to learn that her face was used in a pornographic Deepfake video.
[14] (U) | Britt Paris, Joan Donovan | Data & Society| https://datasociety.net/library/deepfakes-and-cheap-fakes/ | 18 Sep. 2019 | DEEPFAKES AND CHEAP FAKES.
[15] (U) | Raina Davis | HARVARD Kennedy School | Belfer Center| https://www.belfercenter.org/publication/technology-factsheet-deepfakes | Spring 2020 | Technology Factsheet: Deepfakes.
[16] (U) | Britt Paris, Joan Donovan | Data & Society| https://datasociety.net/library/deepfakes-and-cheap-fakes/ | 18 Sep. 2019 | DEEPFAKES AND CHEAP FAKES.
[17] (U) | Britt Paris, Joan Donovan | Data & Society| https://datasociety.net/library/deepfakes-and-cheap-fakes/ | 18 Sep. 2019 | DEEPFAKES AND CHEAP FAKES.
[18] (U) | Chintan Trivedi |Medium| DG AI Research Lab| https://medium.com/deepgamingai/deepfakes-ai-improved-lip-sync-animations-with-wav2lip-b5d4f590dcf | 31 Aug. 2020 | DeepFakes AI- Improved Lip Sync Animations With Wav2Lip.
[19] (U) | K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, C V Jawahar| ACM International Conference on Multimedia| https://doi.org/10.1145/3394171.3413532 | Oct. 2020 | A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild | p 484-492.
[20] (U) | K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, C V Jawahar| ACM International Conference on Multimedia| https://doi.org/10.1145/3394171.3413532 | Oct. 2020 | A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild | p 484-492.
[21] (U) | Shruti Agarwal, Tarek El-Gaaly, Hany Farid, Ser-Nam Lim |IEEE International Workshop on Information Forensics and Security| https://arxiv.org/abs/2004.14491 | 29 Apr. 2020| Detecting Deep-Fake Videos from Appearance and Behavior.

[22] (U) | Hannah Smith, Katherine Mansted |Australian Strategic Policy Institute| https://www.aspi.org.au/report/weaponised-deep-fakes | 29 Apr. 2020 | Weaponised deep fakes.
[23] (U) | Raina Davis | HARVARD Kennedy School | Belfer Center| https://www.belfercenter.org/publication/technology-factsheet-deepfakes | Spring 2020 | Technology Factsheet: Deepfakes.
[24] (U) | Graphika Team |Graphika| https://graphika.com/reports/fake-cluster-boosts-huawei | 29 Jan. 2021 | Fake Cluster Boosts Huawei.
[25] (U)| FireEye | 20 Jul. 2021 | Dual Information Operations Campaigns Promote Lebanese Political Parties Kataeb and Free Patriotic Movement Amid Economic, Political Crisis | A reliable US cyber security company which also releases threat intelligence reports.
[26] (U) | Graphika Team |Graphika| https://graphika.com/reports/step-into-my-parler/ | 01 Oct. 2020 | Step into My Parler.
[27] (U) | Ben Nimmo, C. Shawn Eib, Léa Ronzaud |Graphika| https://graphika.com/reports/operation-naval-gazing/ | 22 Sep. 2020 |Operation Naval Gazing.
[28] (U) | Eto Buziashvili |Medium| DFR Lab| https://medium.com/dfrlab/inauthentic-instagram-accounts-with-synthetic-faces-target-navalny-protests-a6a516395e25 | 28 Jan. 2021 | Inauthentic Instagram accounts with synthetic faces target Navalny protests.
[29] (U) | Henry Ajder, Giorgio Patrini, Francesco Cavalli |Sensity| https://www.medianama.com/wp-content/uploads/Sensity-AutomatingImageAbuse.pdf | Oct. 2020 | Automating Imagine Abuse: Deepfake bots on Telegram.
[30] (U) | Siladitya Ray | Forbes| https://www.forbes.com/sites/siladityaray/2020/10/20/bot-generated-fake-nudes-of-over-100000-women-without-their-knowledge-says-report/?sh=428694037f6b | 20 Oct. 2020 | Bot Generated Fake Nudes of Over 100,000 Women Without Their Knowledge, Says Report.
[31] (U) | Karen Hao | MIT Technology Review| https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/ | 12 Feb. 2021 | Deepfake Porn is Ruining Women's Lives. Now the Law May Finally Ban It.
[32] (U) | Tom Simonite |Wired| https://www.wired.com/story/ai-fueled-dungeon-game-got-much-darker/ | 05 May 2021| It Began as an AI-Fueled Dungeon Game. It Got Much Darker.
[33] (U) | Evan Jacoby |Vice| https://www.vice.com/en/article/vb55p8/i-paid-dollar30-to-create-a-deepfake-porn-of-myself | 09 Dec. 2019 | I Paid $30 to Create a Deepfake Porn of Myself.
[34] (U) | Danielle Keats Citron | Prepared Written Testimony and Statement for the Record before the House Permanent Select Committee on Intelligence | Hearing on "The National Security Challenge of Artificial Intelligence, Manipulated Media, and Deep Fakes" |13 Jun. 2019.
[35] (U) | Jana Benscoter |PennLive| https://www.pennlive.com/news/2021/03/pa-woman-created-deepfake-videos-to-force-rivals-off-daughters-cheerleading-squad-police.html | 12 Mar. 2021 | Pa. woman created 'deepfake' videos to force rivals off daughter's cheerleading squad: police.
[36] (U) | Interview with Karen Hao | 14 May 2021.
[37] (U) | Interview with Karen Hao | 14 May 2021.
[38] (U) | Anne Pechenik Gieseke | Vanderbilt Law Review| https://scholarship.law.vanderbilt.edu/vlr/vol73/iss5/4/ | 05 Oct. 2020 | "The New Weapon of Choice": Law's Current Inability to Properly Address Deepfake Pornography | p 1479-1515.
[39] (U) | Britt Paris, Joan Donovan | Data & Society| https://datasociety.net/library/deepfakes-and-cheap-fakes/ | 18 Sep. 2019 | DEEPFAKES AND CHEAP FAKES.
[40] (U) | Daniella Scott | Elle | https://www.elle.com/uk/life-and-culture/a30748079/deepfake-porn/ | 02 Jun. 2020 |Deepfake Porn Nearly Ruined My Life.
[41] (U) | Anne Pechenik Gieseke | Vanderbilt Law Review| https://scholarship.law.vanderbilt.edu/vlr/vol73/iss5/4/ | 05 Oct. 2020 | "The New Weapon of Choice": Law's Current Inability to Properly Address Deepfake Pornography | p 1479-1515.
[42] (U) | Britt Paris, Joan Donovan | Data & Society| https://datasociety.net/library/deepfakes-and-cheap-fakes/ | 18 Sep. 2019 | DEEPFAKES AND CHEAP FAKES.
[43] (U) | Britt Paris, Joan Donovan | Data & Society| https://datasociety.net/library/deepfakes-and-cheap-fakes/ | 18 Sep. 2019 | DEEPFAKES AND CHEAP FAKES.

[44] (U) | Britt Paris, Joan Donovan | Data & Society| https://datasociety.net/library/deepfakes-and-cheap-fakes/ | 18 Sep. 2019 | DEEPFAKES AND CHEAP FAKES.
[45] (U) | Jon Bateman |Carnegie Endowment for International Peace| https://carnegieendowment.org/2020/07/08/deepfakes-and-synthetic-media-in-financial-system-assessing-threat-scenarios-pub-82237 | 08 Jul. 2020 | Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios.
[46] (U) | Koen Putman, Dario Raffaele, Gerjen van den Dool | Accenture| https://www.accenture.com/nl-en/blogs/insights/deepfakes-how-prepare-your-organization | 06 Oct. 2020 | Deepfakes: How to prepare your organization for a new type of threat.
[47] (U) | Interview with Noelle Martin | 28 May 2021.
[48] (U) | Alex Drozhzhin | Kapersky Blog| https://usa.kaspersky.com/blog/rsa2020-deepfakes-mitigation/21133 | 12 Mar. 2020 | How to mitigate the impact of deepfakes.
[49] (U) | Interview with Dr. Matthew Wright | 18 May 2021.
[50] (U) | Koen Putman, Dario Raffaele, Gerjen van den Dool | Accenture| https://www.accenture.com/nl-en/blogs/insights/deepfakes-how-prepare-your-organization | 06 Oct. 2020 | Deepfakes: How to prepare your organization for a new type of threat.
[51] (U) | Kari Paul | The Guardian| https://www.theguardian.com/us-news/2019/oct/07/california-makes-deepfake-videos-illegal-but-law-may-be-hard-to-enforce | 07 Oct. 2019 | California makes 'deepfake' videos illegal, but law may be hard to enforce.
[52] (U) |David Ruiz | Malwarebytes Lab| https://blog.malwarebytes.com/artificial-intelligence/2020/01/deepfakes-laws-and-proposals-flood-us/ | 23 Jan. 2020 | Deepfakes laws and proposals flood US.
[53] (U) | Lvxiao Chen |The Blog| https://blog.jipel.law.nyu.edu/2020/02/deepfake-is-here-what-should-we-do/ | 14 Feb. 2021| Deepfake is Here. What Should We Do?.
[54] (U) | David Greene | Electronic Frontier Foundation| https://www.eff.org/deeplinks/2018/02/we-dont-need-new-laws-faked-videos-we-already-have-them | 13 Feb. 2018 | We Don't Need New Laws for Faked Videos, We Already Have Them.
[55] (U) | Shruti Agarwal, Tarek El-Gaaly, Hany Farid, Ser-Nam Lim |IEEE International Workshop on Information Forensics and Security| https://arxiv.org/abs/2004.14491 | 29 Apr. 2020| Detecting Deep-Fake Videos from Appearance and Behavior.
[56] (U) | Content Authenticity Initiative |Adobe| https://contentauthenticity.org/ | 2021 | Content Authenticity Initiative Homepage.
[57] (U) | Forbes | https://www.forbes.com/profile/noelle-martin/?sh=6c17ae67306b | 2019 | Noelle Martin 30 Under 30.
[58] (U) | Matthew Ferraro, Benjamin Powell, Jason Chipman |WilmerHale| https://www.wilmerhale.com/en/insights/client-alerts/20200316-fbi-warns-companies-of-almost-certain-threats-from-deepfakes | 16 Mar. 2021 | FBI Warns of "Almost Certain" Threats from Deepfakes.
[59] (U) | Robert Chesney, Danielle Keats Citron |107 California Law Review 1753| https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954 | 21 Jul. 2018 | Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security.
[60] (U) | Cyber Civil Rights Initiative | https://www.cybercivilrights.org/our-services/ | 2021 | Cyber Civil Rights Initiative- What We Do.
[61] (U) | Endtab |https://endtab.org/ | 2021 |Endtab Homepage.
[62] (U) | National Suicide Prevention Lifeline | https://suicidepreventionlifeline.org | 2021 | National Suicide Prevention Lifeline Homepage.
[63] (U) | The Cybersmile Foundation | https://www.cybersmile.org | 2021 | The Cybersmile Foundation Homepage.
[64] (U) | Federal Trade Commission| Identity Theft | https://www.identitytheft.gov/#/ | 2021 | IdentityTheft.gov.
[65] (U) | Without my Consent | https://withoutmyconsent.org/resources/something-can-be-done-guide/evidence-preservation | 2021 | Evidence Preservation.
[66] (U) | Google | https://support.google.com/websearch/answer/9116649?hl=en&ref_topic=9173608 | 2021 | Remove involuntary fake pornography from Google.
[67] (U) | Imatag | https://www.imatag.com/monitor/ | 2021 | How to track images online.

DISCLAIMER STATEMENT: *This document is provided for educational and informational purposes only. The views and opinions expressed in this document do not necessarily state or reflect those of the United States Government or the Public-Private Analytic Exchange Program, and they may not be used for advertising or product endorsement purposes. All judgments and assessments are solely based on unclassified sources and are the product of joint public and private sector efforts.*