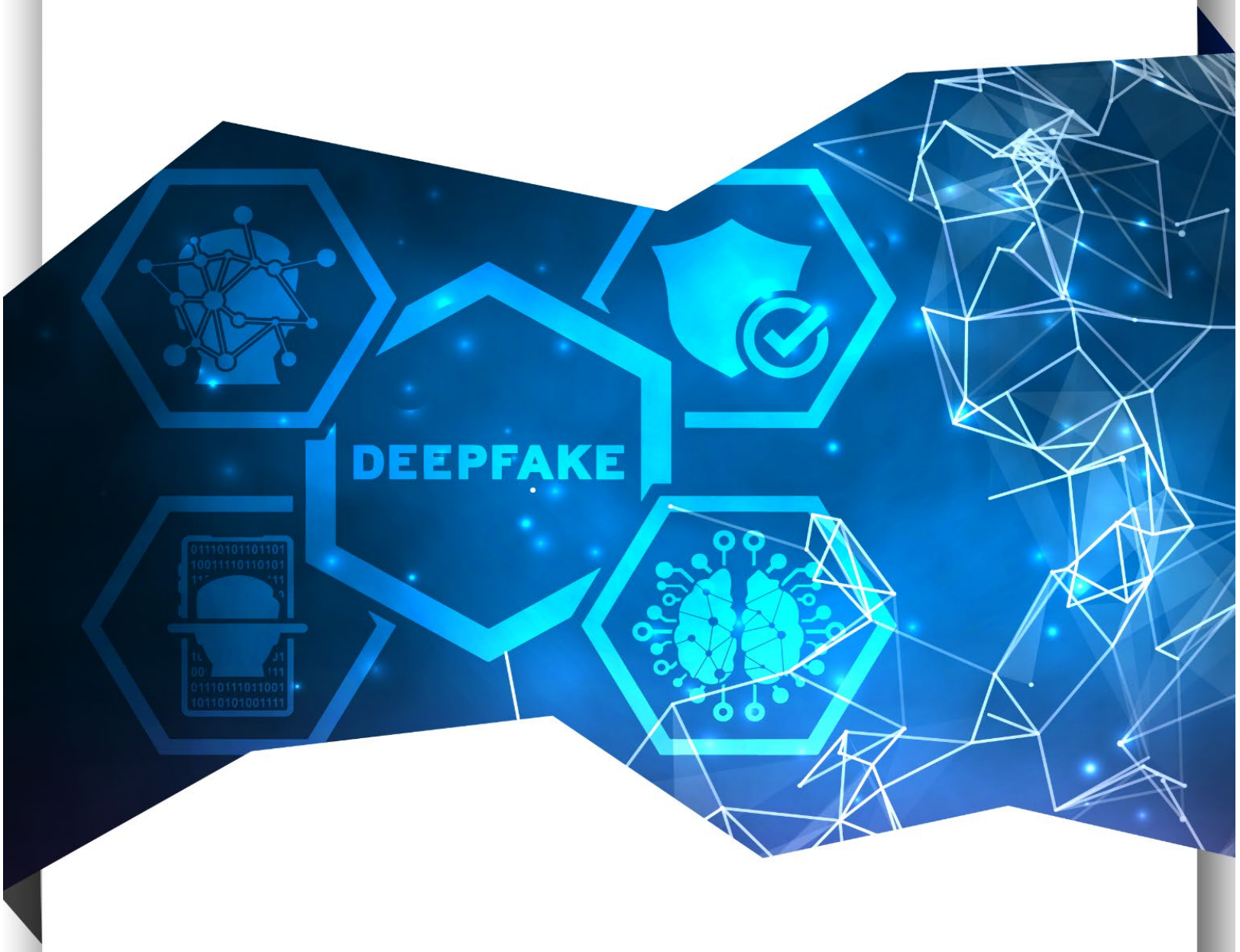
DEEPFAKE

PHRASE TWO || MITIGATION MEASURES



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

Homeland Security

Increasing Threats of Deepfake Identities – Phase 2: Mitigation Measures

Abstract

Deepfakes are a type of synthetic media—commonly generated using artificial intelligence/machine learning (AI/ML)—presenting plausible and realistic videos, pictures, audio or text of events which never happened. In Phase II of our work, we build upon our Phase I findings and offer more in-depth suggestions for organizational, legislative, and regulatory approaches to combat the impending threat of deepfake identities in three use cases. The first use case addresses content offered by creators, owners and immediate users like media organizations, non-government organizations (NGOs), law enforcement and legal institutions that rely on this content. The second use case addresses content disseminated in the broadcast environment where social media platforms and news organizations may be used as vehicles to disseminate false, misleading and ultimately harmful information with broad impacts of varying magnitude. The third use case addresses content associated with real-time or live scenarios for identity proofing and verification to enable and offer services and products. The real-time or near-real-time nature of the interaction in these scenarios make imagery, video and audio content of particular importance. We evaluated these use cases and developed a generalized framework for combatting deepfakes—including an associated checklist—and make recommendations for future work along each of the five aspects of the framework: Establish Policies and Support Legislation; Identify Deepfakes; Demonstrate Integrity, Authenticity, and Provenance; Act Appropriately; and Engineer the Environment.

TEAM INTRODUCTIONS

Tina Brooks, Verizon
C. Pauline Daniel, New York Medical College & United Nations Population Fund
Jesse Heatley, JP Morgan Chase & Co.
Scott Kim, Experian
Maureen R., U.S. Department of State
Burak Sahin, Deloitte & Touche
James S., U.S. Department of Homeland Security
Oliver T., Federal Bureau of Investigation
Richard V., Federal Bureau of Investigation (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."

Introduction

In 2021, our Team summarized the challenges that deepfakes, synthetic media, and disinformation pose to our society in a report titled "Increasing Threat of Deepfake Identities."[1] We noted that these challenges can impact individuals and institutions from small businesses to nation states, but that there may be opportunities to mitigate the threat.

This report discusses some of those mitigations in the context of three specific use cases: (1) Organizations which seek to **actively demonstrate the authenticity** of recorded content they have created; (2) Organizations which **broadcast** recorded content created by others; and (3) Organizations engaged in **"real-time" communications** with individuals, wherein the organization seeks to verify that the individual is not a synthetic representation.

Organizations to whom the first environment applies could include news media, law enforcement, equipment manufacturers, and non-government organizations (NGOs) which seek to present factual representations of events, because their success depends on others (i.e., the public or their customers) accepting the reality of what is being presented. The second environment includes the news media, but also includes social media service providers and other providers of third-party content. In this second environment, there may or may not be a demand from the recipients of the content (e.g., "customers") to have its authenticity provided. Finally, organizations operating in the third environment could include both commercial and government entities which seek to provide services to customers, but have a need to verify that the customers are real people in order to avoid fraud.

While the first two environments describe the delivery of content to a customer, the third environment would have the customer delivering the content. The degree to which that content is accepted by the recipient will depend on deepfake mitigation measures incorporated into the process, which we describe below. Regardless of the environment, however, the ultimate decision regarding whether to accept the content as real or not will come down to individual human beings, who will need to be prepared.

Organization of this Report

This report is organized as follows:

- We first describe three broad use cases, noting individuals or organizations which may be affected and the primary challenge deepfakes pose within that use case.
- We then offer a general framework for combatting deepfakes with several elements, whose importance will vary with use case.
- Finally, we identify ways in which those elements may be implemented today, or in the future.
- Several appendices are included to provide the interested reader with more insight into some of the specific concepts described herein.

[1] Public-Private Analysis Exchange Program, "Increasing Threat of Deepfake Identities," https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf, 2021.

Use Case #1: Content Creators/Owners

The malicious misuse of synthetic content and deepfakes pose a threat to any company, organization, or government entity that relies on public - or a customer's - trust to achieve its mission. When seeing is no longer believing, trust in companies, non-government organizations, law enforcement agencies, and the legal system erodes, facilitating an inherently unstable and distrusting environment. Bad-faith actors will weaponize this environment of distrust to further polarize an already fractured public. Instead of fair scrutiny, original content creators may face pessimistic doubt from an audience weary of, or susceptible to, deepfakes. While this paper critically looks at the implications within the judicial system, we acknowledge other classes of content creators who have a need to ensure their content is legitimate and authentic, such as:

- A journalist investigates a refugee camp and reports on the conditions through photographs and recorded interviews with refugees.
- An NGO collecting satellite imagery to report on ecological changes to an area.
- Several individuals using their phones to record an incident involving police attempting to detain an individual.

Use Case: During a child custody case in Britain in 2019, one of the parties introduced an audio recording of her former husband being verbally abusive to their child in an effort to get full custody.[1] The former husband admitted that the recording sounded like him, but he was adamant that the recording was not actually him. "This is always a difficult position to be in as a lawyer, where you put corroborating contrary evidence to your client and ask them if they would like to comment. My client remained, however, adamant that it was not him despite him agreeing it sounded precisely like him, using words he might otherwise use, with his intonations and accent unmistakably him. Was my client simply lying?"

The Need to Verify the Integrity and Authenticity of Evidence

Although news media organizations, academia, and other institutions have a necessary and compelling interest to ensure their content is real and verifiable, the loss of content authenticity could have far greater consequences in the realm of law. Strict and specific rules and regulations govern how evidence can be admitted and presented into the courtroom. The mere existence of synthetic content and deepfakes may introduce doubt into court proceedings. This applies particularly in criminal cases when justice and an individual's freedom are at stake, and even the smallest sliver of doubt can undermine a case. Without proper evidence handling by law enforcement agents, strict adherence to the chain of custody of evidentiary information, advanced, collaborative digital forensic analysis, and judicial education, deepfakes could threaten the trust and stability of the judicial system.

One of the underlying principles of American democracy is founded in due process of our judicial system to guarantee the right of a fair trial. The court presumes that digital images, videos, body-worn camera footage, and security camera footage presented in court are

authentic and real. While media forensic experts regularly testify, we may soon need experts attuned to synthetic media technologies to testify, as well. Here are the core stakeholders:

- **Law Enforcement Agencies (LEAs)**: including local, state, tribal, and federal law enforcement agencies who are responsible for investigating crimes, documenting, and safeguarding evidence. Deepfakes could be collected as evidence and unwittingly passed on to the prosecution. Additionally, it could be possible for a malevolent actor to hack into the digital storage of evidence and corrupt data or add deepfakes.

- **Lawyers**: Both the prosecution and defense are responsible for ensuring evidence presented in a case was legally acquired, is accurate, and has not been tampered with in any way.

- **Judges**: Among other duties, judges are responsible for determining whether evidence submitted by the prosecution or defense is admissible.

- **Jury:** In a criminal trial, a juror's job is to determine the innocence or guilt of the defendant; in a civil trial jurors must determine whether the defendant is at fault for the offense and to what extent. The reverse CSI effect[2] refers to a jury's belief that any piece of digital evidence, no matter how convincing, could be a deepfake. Therefore, without ironclad proof that a piece of evidence is not a deepfake, the evidence is less trustworthy. This would put the burden on lawyers to prove that the video and audio evidence they are relying on is real. In order to carry out their job, the jury may need to receive specialized education and training to understand the basics of deepfakes.

- **Forensics Experts**: Although media and other forensics experts routinely testify in criminal and civil cases and many of them may well have expert knowledge on AI-enabled synthetic content, deepfakes, and associated technologies, the rapid development of content generation, detection, and authentication may require experts to continuously evolve their understanding of this issue.

Use Case #2: Broadcast Environment

Use Case: In March 2022, social media platforms circulated a breaking news video of Ukrainian President Zelenskyy urging soldiers to surrender in their fight against Russia (See Figure 1). Speaking behind a podium, Zelenskyy admitted defeat, acknowledging Ukraine had "decided to return Donbas" and that the nation's war efforts had failed. The video message, however, was fabricated–likely by Russian disinformation agencies.[2] That said, if accepted as truth, the fast-spreading video could have impacted the lives of millions and, potentially, affected the war more than traditional armaments and troops.

[2] The CSI effect refers to "the prosecutors' belief that crime programs are skewing jurors' courtroom expectations … they want the smoking gun and the DNA proof - the unmistakable proof that points to the perpetrator."

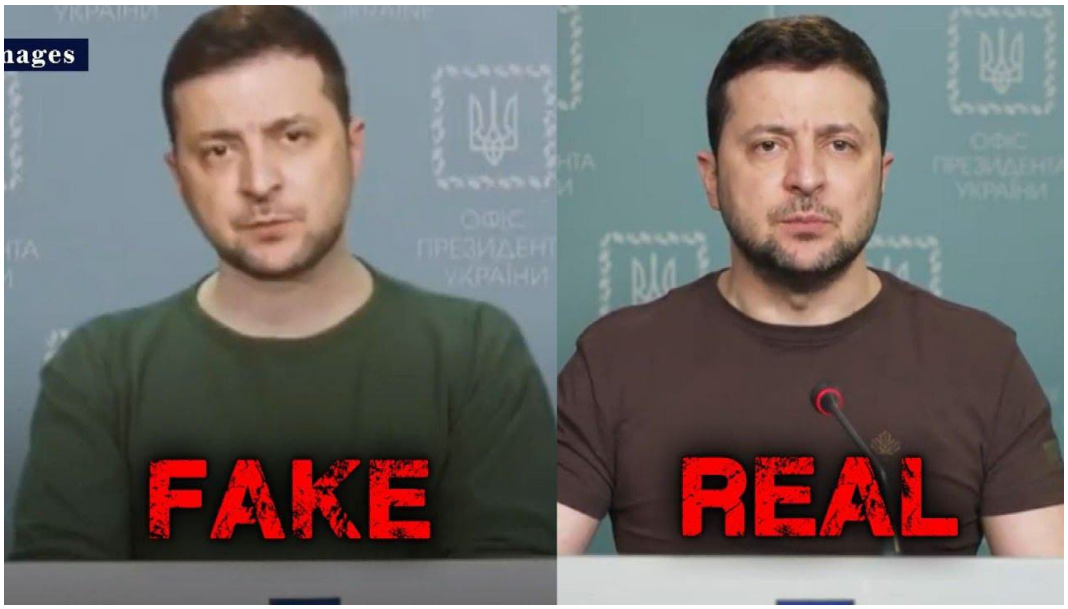


Figure 1. Screenshot from Zelenskyy deepfake video compared to a real picture of the Ukrainian President.[3]

With global conflicts, financial systems, and other fast-moving sectors, there is little doubt that deepfakes pose a growing threat to society, especially with media companies' ability to reach to billions worldwide. As deepfakes continue to develop technically, through artificial intelligence, it will only become easier to fool some of the people or organizations even some of the time, all which could have disastrous consequences. This may not be too far off. It has clearly become easier to mimic real people in what appears to be authentic videos and these videos, if accepted as truth like in the Ukraine scenario, could greatly impact geopolitical emergencies, political crises, and other world events. That said, our research and discussions have shown that many hopeful efforts are now taking shape in the broadcast media and tech space. Of course, future efforts to address manipulated content will likely be driven by consumer demands, company financial concerns, and regulatory oversight.

The Need to Identify and Act on Deepfakes in Broadcast Media

In recent years, even lower barriers to the development of high-quality deepfakes has led to a surge in misinformation campaigns or realistic news content.[4,5,6] Although traditional newsrooms may have always wrestled with some fact checking or layers of verification, new deepfakes may create new challenges for news operations, especially smaller-scale companies, and create new concerns over trust in the news media and journalism. The evolving deepfake technologies likely present several significant dangers to news media institutions and professionals, which create the need to identify and act against deepfakes. This challenge extends beyond traditional news media institutions, to include social media companies, which now serve as a primary source of news and information for many in the world. Certainly, the broadcast world will not want to face reputational risk, commercial losses, legal risks, or government regulatory scrutiny for failing to address the threats:

- *Reputational Risk*: Believable deepfakes are making it easier to damage broadcast companies' reputations; and this risk will become harder to manage and mitigate as the underlying technologies continue to evolve.

- *Commercial Risk*: News and social media rely on fast-paced delivery of news and information that compounds the risk associated with deepfakes. The unintentional distribution of falsified media can have far-reaching effects including damaging reputations and impacting consumer behavior, as well as affecting a company's stock prices.

- *Legal Risks and Liabilities*: Deepfakes' impact could create tricky legal challenges and liabilities. In the political arena, a fake video can shift the tide of a tight election if it is released the day before a vote and cause civic unrest. In the corporate arena, a deepfake can cause a company's stock to move up or down erratically, affecting global markets.

- *Political Backlash and Regulatory Concerns*: As China and other countries are issuing new rules to clamp down on deepfakes, targeting the distributors not the creators, broadcasters are likely fearful that increased regulatory scrutiny could add pressure on them.

Use Case #3: Real-Time Verification

In an identity-based transaction that enables logical or physical access for credentialed or registered individuals, both the individual on whose behalf the identity assertion is made and the organization verifying the assertion have a vested interest in the accuracy of the assertion. Fraud can result in funds being distributed to the wrong parties and have immediate remunerable consequences, but it can also have direct adverse impacts on an organization's brand, trustworthiness, and reputation. Moreover, it is possible that new forms of liability and insurance connected to identity risk may emerge in the future, further complicating the risk organizations assume in identity transactions.

The key directive of identity transactions is to ensure that the person claiming an identity is the actual person to whom the claimed identity belongs. An example to illustrate this use case could involve deepfake voice phishing (vishing), which uses cloned voices to impersonate trusted individuals over the phone.[7]

Use Case: Let us consider a scenario wherein a media company has experienced several years of financial difficulty and is in the process of getting purchased by a wealthy investor. The CEO joins the latest board meeting by phone instead of in person due to a scheduling conflict, whereas all other members are present at company headquarters. His identity is not authenticated during the phone call. As the other board members discuss the offer and whether the investor is a good fit for company, the CEO states the investor is not a good fit, bringing up a safari the investor went on, killing an endangered elephant. The CEO suggests animal activists would protest the company and the brand would be severely damaged. He

concludes that the brand would not recover and the company would not survive. After the board members consider this, they take a vote and reject the investor's offer.

The next day the CEO received an email thanking him for sharing the investor's cruel actions. The CEO called the board member immediately to report that he did not attend the meeting. The board member disputes this, recalling the conversation and the detail about the endangered elephant, mentioning the CEO's passion about the matter. The CEO asks if the investor has been notified that his offer is rejected, and indeed he has been. As a result, the incident is made public and the company's brand suffers significant damage, with many media articles questioning how such a tech giant could let this happen. In addition, users of the platform have concerns about their own information being secure.

The Need to Verify Identity in Real-Time Interactions

Identity verification is a key component of Identity, Credential and Access Management (ICAM) systems[3,8] and modern systems typically verify identities using multiple factors in a methodology commonly referred to as multi-factor authentication, or MFA. The composite representation of an identity using multiple factors is statistically stronger than any one of the factors alone and organizations implementing MFA reduce identity risk by strengthening their authentication methods. As an example, a four-letter PIN would offer a false accept rate of approximately 0.01% alone, i.e. 1 in 9999 attempts has a chance of providing access to an unauthorized user. If the ICAM system using the PIN as an authentication factor, however, were to add face biometric verification as a second factor and utilize a face recognition algorithm that yields no more than 5 false positives out of 100,000 matches, the combination of two factors can bring the false acceptance rate down to 0.00005%.

Authentication factors come in three categories:

- **What one knows:** Factors the person demonstrates knowledge of, such as a password or a PIN
- **What one has:** Factors the person is in possession of and provides, such as a driver's license
- **What one is:** Factors intrinsic to the person's identity, such as their biometrics

While biometrics have traditionally been considered as harder to compromise (not to mention convenient in comparison to multiple passwords or more secure than physical identity tokens), an increasing acceptance of non-proctored authentication use cases has also broadened the threat landscape, leading organizations to integrate various fraud detection and mitigation measures. The interest in contactless and remote transactions, which was growing prior to the pandemic, has only accelerated with Covid-19.

[3] According to the General Services Administration, Identity, Credential and Access Management (ICAM) comprises the tools, policies and systems that allow an organization to manage, monitor and secure access to protected resources.

When the factor of authentication in a real-time verification transaction includes the presentation of an image, video, or audio, as is the case in biometrics, the threats to the security of the transaction include more than the baseline cyber security risks that underlie all computer systems and components. As depicted in the flow diagram below (Figure 2)[9], while each system component and the transmission of data between each component adds to the vulnerability of the system, the biometric nature of the system introduces additional vulnerabilities.

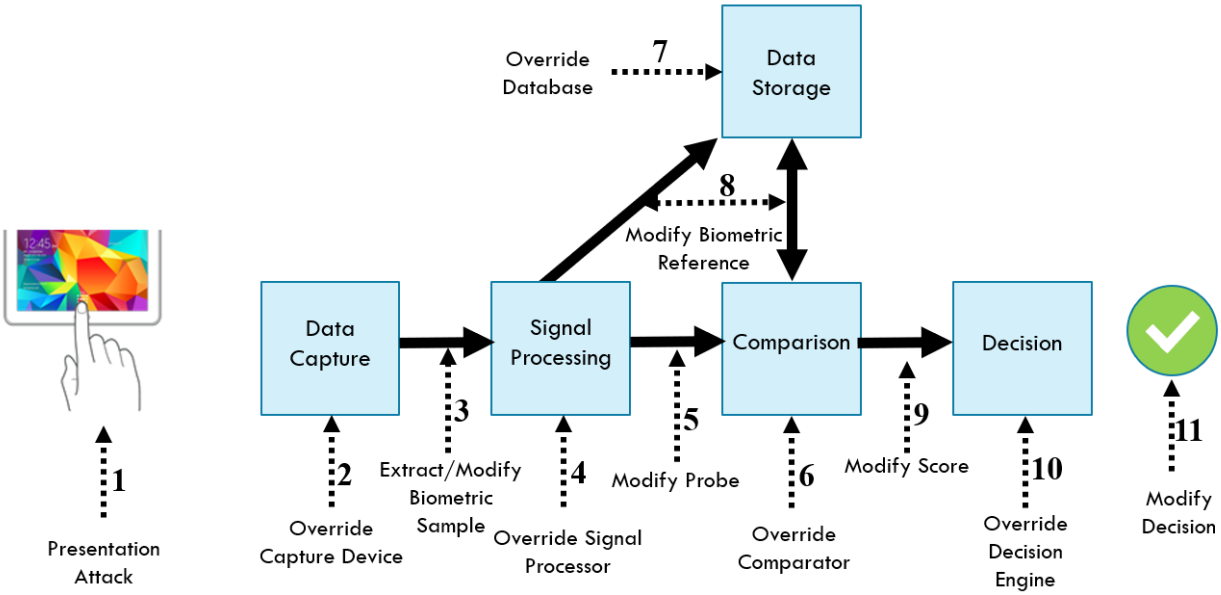


Figure 2. Authentication Process Flow.

According to ISO/IEC 30107-3:2017, the international standard on biometric presentation attack detection, a presentation attack is the presentation of "an artifact or of human characteristics to a biometric capture subsystem in a fashion intended to interfere with system policy."[10] Presentation attacks can be in the form of gummy fingers constructed with an impression of someone's fingerprints, an audio recording submitted to verify someone's voice, or a face video of an account holder replayed to gain access. In the case of face or voice recordings, these can be genuine recordings of the actual person captured (or downloaded from the public internet) by an attacker, or recordings generated synthetically using the likeness of the person. The latter, deepfakes, are becoming an increasingly concerning threat and the current state of the technology does not offer a lot of options to detect them.

Traditional presentation attacks are better known and more easily developed, and attackers fall on them more easily, even though arguably generating a deepfake video that can synthesize the victim saying a specific sentence or performing a challenge-response action or mimic can be easier than creating these recordings only using existing material belonging to the victim.

A Generalized Framework for Combatting Deepfakes

Given the dangers of deepfakes, organizations both public and private are adapting to these threats. Our review of actions against deepfakes and manipulated media in general suggests a framework to combat deepfakes and support best practices (See Appendix A for a graphic representing the Framework, Appendix B for a checklist to assess your preparedness, and Appendix C for current actions aligned to these steps):

1. Establish *policies* and support *legislation* that allow organizations to scrutinize media and act appropriately when necessary
2. Develop the capability to *identify* deepfakes
3. Advance ways to demonstrate the integrity, authenticity and provenance of media
4. *Act appropriately* when problematic deepfakes are detected
5. Proactively *engineer the environment* to reduce the effects of deepfakes.

Establish Policies and Support Legislation

The establishment of policies and legislation to provide the basis for scrutinizing media and acting appropriately when necessary is the first line of defense against deepfakes. Social media platforms, especially the large ones such as Meta, YouTube, and Twitter, have been proactive in establishing policies against synthetic and manipulated media. Government has also seen the need to take action and have enacted legislation against deepfakes.

Identify Deepfakes

Mitigating the effects of a deepfake is more effective if you are able to identify and separate it from the sea of "regular" media. Media companies engage in efforts to identify inappropriate content before it is posted, as well as, flag content after it is on their site. This identification effort includes technological advances, use of content and review teams, and partnering with users and other third-parties. Solving for timely identification of deepfakes will be key in stemming the most serious threats.

Demonstrate Integrity, Authenticity, and Provenance[4,11,12]

Existing measures to demonstrate integrity and authenticity vary in technique and effectiveness, and some have universal acceptance as a standard adopted across different

[4] For the purpose of this paper, we define integrity in relation to the process of "integrity verification" which is defined in ASTM E1732-19 ("Standard Terminology Relating to Forensic Science") as "a process of confirming that the data presented is complete and unaltered since time of acquisition." Authenticity primarily means to validate that the content is legitimate. Provenance primarily establishes technical documentation of the origins of content, its history to include any alterations, and chain of custody. For example, a journalist video records an interview with a politician. The recording is uploaded and shared on Twitter. Integrity means that the video file is forensically unaltered. Authenticity means that the video is verified to be the actual recording of the actual interview with the politician. Provenance could mean that the hardware used to create the original video file creates metadata that details the origins, technical, and record keeping information of the file, any changes, where it was disseminated, and so forth.

domains, applications, and use cases. Advances in technology have evolved some of these standards, as cryptography, computer science, and digital forensic analysis have improved. (See Appendix D for more on current efforts in ICAM technologies)

Act Appropriately

Companies and organizations must address the challenge of appropriate responses. Once a company or organization has identified a deepfake, there is the challenge of what action is necessary. Should all manipulated media be deleted or removed? What actions should the company or organization take against the media, channels, individuals who posted the content, and/or content creators? How do they decide what to do?

Engineer the Environment

Companies and organizations can *engineer the environment* to promote the truth and reduce the effects of deepfakes. Creators can proactively embed authentication measures into their content. Organizations utilizing real time interactions, can literally engineer the environment in ways that can defeat attempts to use deepfakes. Broadcasters have multiple options for action, such as the support of authoritative voices, education of the public and internal workforce, and the building of collaborative relationships. Companies and organizations can do more than react to manipulated content when it arises. They can take proactive actions to reduce risk and build resiliency and capabilities before a crisis event occurs.

Future Efforts to Combat Deepfakes

It is clear that public and private organizations need far more robust measures and strategies to combat deepfake technologies. That said, our research and discussions have shown that many hopeful efforts are now taking shape. Future efforts to limit deepfake influence will coalesce around the social, economic and political pressures. As we have seen earlier, the motivations to address to deepfake technologies in the broadcast environment are manifold—from consumer pushback to reputational damage. The likely motivations and the future efforts will be far different for public and private organizations. For instance, we envision efforts will require the technology competency or standardization of leading U.S. government departments and agencies, as well the efficient outreach and delivery of private companies. We assess that future efforts against deepfakes will vary based on the relative strengths and weaknesses of private and public organizations:

Establish Policies and Support Legislation

- **Public Policy and Legislative Actions:** As the U.S. government seeks to counter deepfake technology, the U.S. Congress may take actions to pass legislation aimed at creating new standards, tools, and leadership around counteracting deepfake

threats. In 2021, Senators Rob Portman (R-OH) and Gary Peters (D-MI) introduced the Deepfake Task Force Act, which includes tech and media participants. That legislation, yet to be passed, received strong support from technology leaders and industry. It will likely either move forward or form the basis for future attempts to address the problem of altered digital content in the broadcast space.

- New Government Agencies or Departments or New Requirements: In response to the growing threat of misinformation, the U.S. government could seek to establish new federal agencies or organizations to identify manipulated content threats, share information, and assist with alerting broadcast media companies, public broadcasting organizations, and audiences about false and misleading content. Similarly, in 2018, the U.S. Congress created the Cybersecurity and Infrastructure Security Agency (CISA) under the Department of Homeland Security to address cybersecurity threats to critical infrastructure. Alternately, new requirements could be levied on existing agencies to address these challenges themselves.

  First Amendment Considerations: In the United States, any legislative restrictions against the creation of synthetic content and deepfakes will likely face First Amendment challenges.[13] The First Amendment protection of expression is unique in its scope, integrity, and strength.[14] Should Congress or the courts find it necessary to take action against the misuse of synthetic content, society will still need to contend with those actors undeterred by any legislative and/or judicial restriction. Deepfakes present a different take on the eternal struggle of balancing the interests of the First Amendment and the public good. As technology advances, Congress and the Supreme Court must decide how to reconcile fundamental constitutional freedoms.[15]

Identify Deepfakes

- New Technical Standards & Data Labeling: The U.S. government will likely continue to move forward with plans to create more standards for commercial enterprises, including news and social media, for the detection and mitigation of deepfake technologies and adversarial content. As with Identifying Outputs of Generative Adversarial Networks (IOGAN) Act (H.R. 4355), the U.S. Congress seeks to direct national agencies to conduct research and develop measurements and standards on deepfake technologies.

- New Detection Technologies: Future efforts will continue to emerge to address the technical needs related to deepfake technology production and detection. We expect to see more high-profile initiatives, like the Deepfake Detection Challenge (DFDC), an open, collaborative initiative to produce innovative new technologies to detect deepfakes and manipulated media. Launched in December 2020, the DFDC brought together 2,114 participants who submitted more than 35,000 models to the competition.[16]

- New Entrants into Deepfake Defense Technologies: In the past few years, several startups have emerged to tackle deepfake threats with innovative solutions to

counter fake content. Given the financial and legal motivations outlined above, we assess that there is growing space for technology startups and service providers offering deepfake identification platforms. As these detection and mitigation technologies become more prevalent, it may lower service fees for a broader range of broadcast and media clientele.

- **Developing and Utilizing Large Training Data Sets:** There is increasing industry interest in seeking guidance and help from the National Institute of Standards and Technology (NIST) in developing and utilizing large data sets to train systems and improve performance. As an example, Paravision, provider of one of the leading face biometric technologies, is currently developing deepfake detection algorithms in response to a request from a U.S. government agency and a foreign partner intelligence agency.[17] Leaders from the company have acknowledged that a crucial part of the development process involves generating large, labeled data sets to train and test algorithms.

- **Using Identifying Factors Accompanying the Transmitted Data:** Forensic methods that do not attempt to detect the deepfakes from the transmitted (and modified) signals but instead consider all the identifying factors accompanying the transmitted data (such as a verified source, digital signatures, etc.) are also under consideration against deepfakes, and utilizing these methods as yet another authentication factor that contributes to the overall strength and assurance of an integrated system is the approach we expect to continue.

Demonstrate Integrity and Authenticity

Standardized Data Labeling & Attribution: In the years ahead, broadcast and media content may need to adopt innovative labeling or attribution standards. The Content Authenticity Initiative (CAI) and Coalition for Content Provenance and Authenticity (C2PA) are working to create new labeling standards so digital content, such as news media, photography, digital art, and video recordings cannot be secretly altered. Perhaps the keystone element in establishing trust in the origins and life cycle of digital content, this effort endeavors to implement a technical standard that is ubiquitously interfaceable with content creators, consumers, broadcasters, hardware, and software. A transparent audit log will reveal any content edits or alterations in a cryptographically secure fashion, providing viewers with the knowledge of if the content they consumed was modified in any way since it was originally created. Figure 3 shows how content leveraging CAI standards would establish digital provenance through every step of the process from creation to viewing.

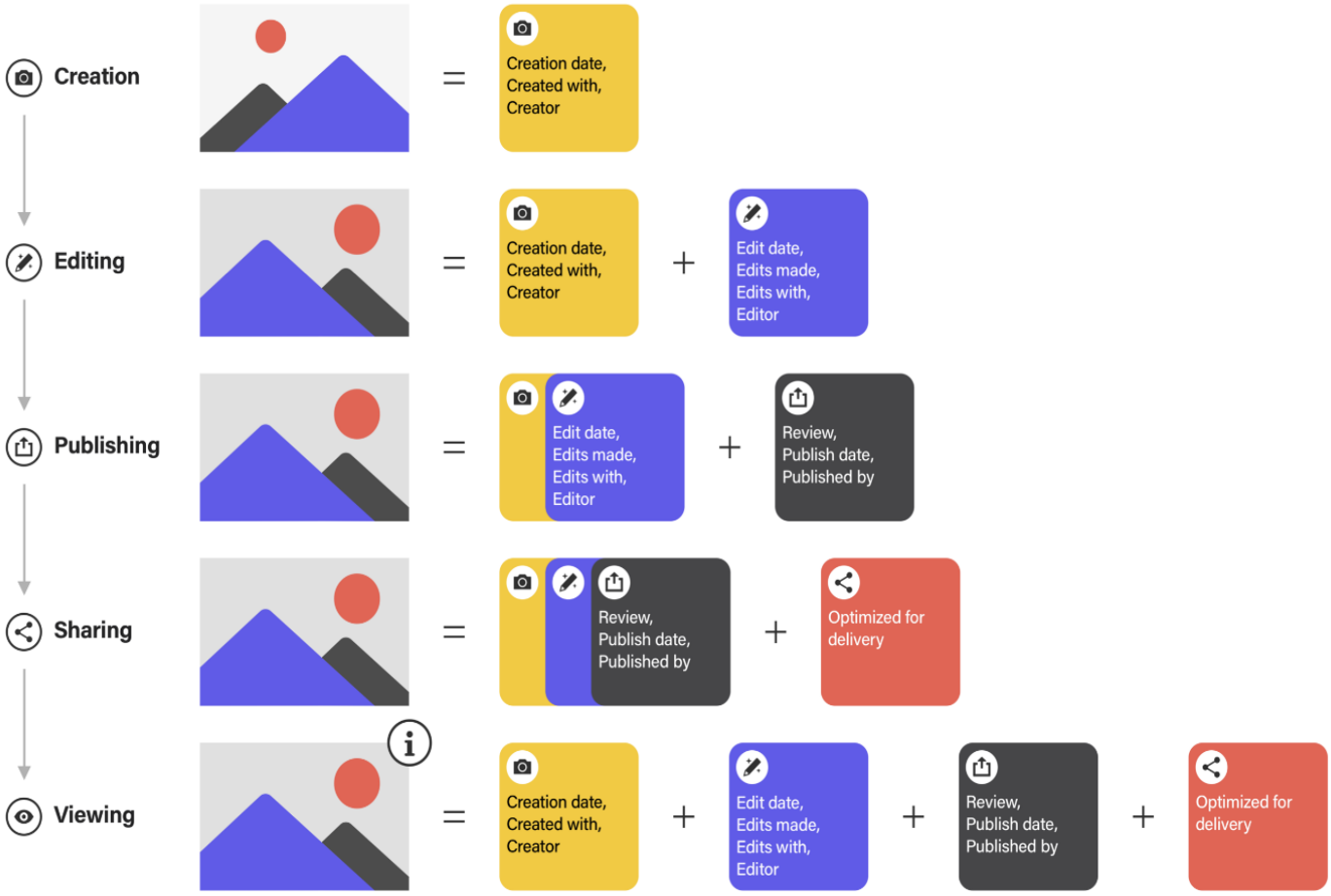


Figure 3. Provenance-supporting signatures can be embedded in digital content at each step in the process from content creation to viewing, including in multiple editing steps. (Graphic provided courtesy of Adobe, Inc.)

- **Perceptual image hashing:** This is a newer technique that incorporates machine learning to achieve higher success of image identification and retrieval through similarity analysis including duplicate, near duplicate, and similar images. This method also achieves higher confidence in authenticating images and tampering detection in many cases including against digital watermarking, noise addition, and contrast adjustment or scaling.[18] Instead of assigning enciphered text to label an image purely based on the underlying bit data, perceptual hashing can assign the same hash value for similar images. Deep perceptual hashing methods, which incorporate deep learning and perceptual hashing, enhance accuracy and speed of image detection within large databases, and improve image authentication and tamper detection.

- **Blockchain Authentication and Data Integrity:** Newer technologies such as blockchain could also provide some utility in authenticating images and data integrity. Recent research had demonstrated proof of concept of a blockchain-based system integrating several different algorithms for image authentication with mixed success against modifications such as blur, crop, rotation, and image flip.[19] The primary advantage blockchain frameworks could have is the lack of a central authority in the system, which in theory eliminates a possible point of failure.

- **Labeling and Referencing against Databases:** Images that are deemed to be inauthentic synthetic content could be labeled and assigned a hash value which could be cross referenced against a proposed database of existing deepfake content. This strategy could help mitigate against resilient deepfake content that sustains a presence across social media platforms. These techniques can also be useful in identifying content that has been manipulated in some way but remains saliently similar to the original content.

Act Appropriately

- **Effective Labeling of Manipulated Media:** The Partnership on AI and First Draft have come up with a set of twelve design principles for labeling manipulated media that include: don't attract attention to the mis/disinformation; make labels noticeable and easy to process; encourage emotional deliberation and skepticism; offer flexible access to more information; use a consistent labeling system across contexts; repeat the facts, not the falsehoods, etc.[20] More research could be done to determine the most effective ways to label manipulated media and protect the public.

Engineer the Environment

- **Cross-Industry Partnerships:** Several industry partnerships are leading the way to stem deepfakes in media, including C2PA and CAI. C2PA, which includes members Adobe, Microsoft, Arm, Intel TruePic, and the BBC, released its first version of its technical specification for digital provenance in early 2022. Announced by Adobe in 2019, CAI was formed in partnership with Twitter and the New York Times—and it clearly looks to create new opportunities for the media space.

- **New Educational and Training Support:** As the U.S. government has moved to support cybersecurity in leading engineering and computer science programs, the U.S. government could work with industry to create new training materials and educational requirements for journalism, technology, and other impacted industries. Despite the technical aspects, many experts say current deepfakes rarely outsmart human intuition, due to slight nuances like eye blinks and gestures; this suggests proper training could thwart some deepfake threats. (See Appendix E for more on education and awareness efforts)

    - **Law enforcement Education and Awareness:** Educate relevant personnel including patrol officers, evidence, technicians, investigators, and command staff. Include a basic overview of what synthetic content and deepfakes are, how they are made, examples of their usage and methods to detect them. Integrate this type of training into already existing training such as digital forensics, chain of custody, and best practices for protecting digital evidence, including body worn camera footage, from tampering.

    - **Juror Education to Preempt the "Reverse CSI effect"** – It may be necessary to consider bolstering juror instruction, to augment the already existing forensic

experts who routinely testify, to account for the proliferation of synthetic content and deepfakes that may inevitably come to the courtroom. This recommendation comes with the following caveat: It is possible that by educating jurors about deepfakes and the reverse CSI effect, legal officers could end up creating the problem they are trying to preempt. This recommendation requires further study before being operationalized.

- Increase Federal-State-Local-Tribal Collaboration: Individual police departments do not necessarily have the time, space, resources, expertise, or personnel to devote to in-depth digital forensics work. As technology advances, it can safely be assumed that a growing majority of cases will include digital evidence and that the amount of evidence per case will also increase.

    - Increase the size and scope of programs like the Regional Computer Forensic Laboratory program: There are currently 17 Regional Computer Forensic Laboratories (RCFLs) run by the Federal Bureau of Investigation in partnership with other federal, state, and local enforcement agencies. According to the RCFL website, the "laboratories provide forensic services and expertise to support law enforcement agencies in collecting and examining digital evidence for a wide range of investigations, including child pornography, terrorism, violent crime, and fraud."[21] RCFLs or comparable entities could be ideally situated to deal with deepfakes in digital forensics.

- Public-Private Technology Cooperation & Research Support: The U.S. government has the technology competency and resources to advance technical detection and support for the broadcast environment, particularly smaller players, through greater cooperation and research support. The U.S. government currently supports grants and technology integration services programs for specific programs and initiatives. It could create new projects and directives to encompass these threats to media and broadcast entities, in the public and private spheres.

- Partnerships and Competency Assessments: To combat deepfakes, the relative competencies and deficits of public and private institutions could be assessed and form the basis for partnerships and collaboration. The strengths of one organization could help improve the defenses of another. The U.S. government will likely continue to pursue valuable public-private partnerships to boost at risk media and broadcasting organizations from the deepfake threat.

What next?

Regardless of the environment, the ability of individuals and organizations to mitigate the threat posed by deepfake identities will depend upon a partnership between those directly responsible for producing and delivering multimedia content and those consuming it. It is the responsibility of all those engaged with such content to understand their options for verifying the message that content delivers. Individuals and organizations should take steps to ensure that they and their personnel are aware of the risks and resources available to address those risks.

ACKNOWLEDGEMENTS

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

Vladimir Barash, Chief Scientist, Graphika

Kevin Bowyer, Schubmehl-Prein Family Professor, Computer Science Engineering, University of Notre Dame

Edward Delp, Charles William Harrison Distinguished Professor of Electrical and Computer Engineering and Professor of Biomedical Engineering, Purdue University

Amruta Deshpande, Senior Research Scientist, Graphika

Wendy Dinova-Wimmer, Senior Digital Media Architect, Adobe

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

Amyn J., United States Government

Siwei Lyu, SUNY Empire Innovation Professor, University at Buffalo

Jennifer Mathieu, Chief Technology Officer, Graphika

Andy Parsons, Senior Director Content Authenticity Initiative, Adobe

Jonathan P., Deputy Director Biometric and Identity Technology Center,Technology Centers Division, Office of Science and Engineering, DHS Science & Technology Directorate

Mert Sabuncu, Associate Professor, Cornell University

Kate Saenko, Director Computer Vision and Learning Group, Boston University

Jana Schwartz, Principal Researcher for Intelligence Tools & Applications, STR

Neil Serebryany, CEO, Calypso AI

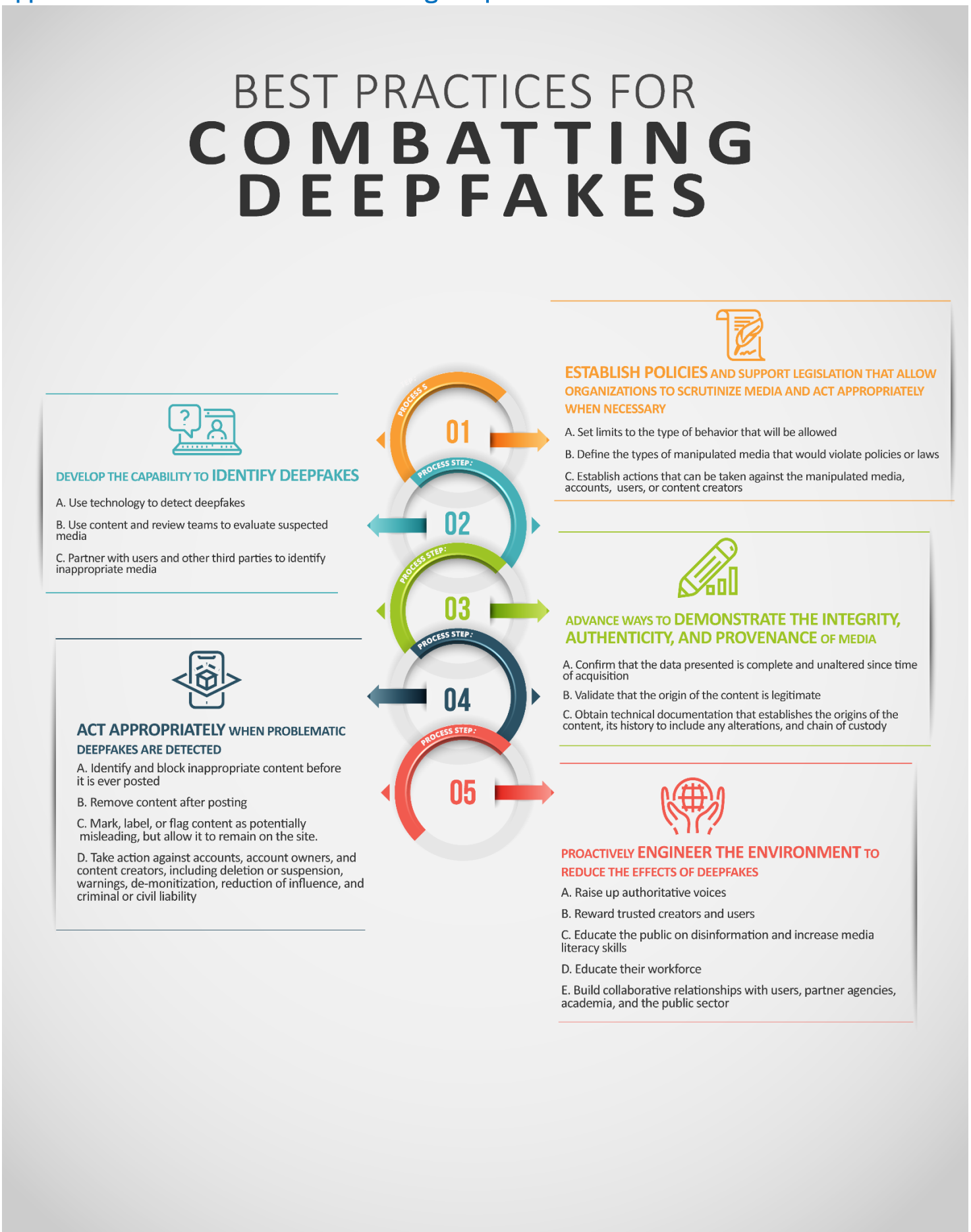
Kirill Trapeznikov, Principal Scientist, STR

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

Annalisa Verdoliva, Associate Professor, University Federico II

Tyler Williams, Director of Investigations, Graphika

Appendix A. Best Practices for Combatting Deepfakes



Appendix B. Checklist for Assessing Deepfake Preparation

HOW PREPARED ARE YOU FOR A DEEPFAKE ATTACK?

		NO	NOT SURE	YES	N/A
1.	HAVE YOU ESTABLISHED *POLICIES* AND/OR SUPPORTED LEGISLATION THAT ALLOW YOU TO SCRUTINIZE MEDIA AND ACT APPROPRIATELY WHEN NECESSARY?	◎	◎	◎	◎
	Do your policies set limits to the type of behavior that will be allowed?	◎	◎	◎	◎
	Do your policies define the types of manipulated media that would violate your policies?	◎	◎	◎	◎
	Do your policies define the types of manipulated media that would violate your policies?	◎	◎	◎	◎
2.	HAVE YOU DEVELOPED THE CAPABILITY TO IDENTIFY DEEPFAKES?	◎	◎	◎	◎
	Are you using technology to detect deepfakes?	◎	◎	◎	◎
	Are you using content and review teams to evaluate suspected media?	◎	◎	◎	◎
	Are you partnering with users and other third parties to identify inappropriate media?	◎	◎	◎	◎
3.	ARE YOU ADVANCING WAYS TO DEMONSTRATE THE INTEGRITY, AUTHENTICITY, AND PROVENANCE OF MEDIA?	◎	◎	◎	◎
	Can you confirm that the data presented is complete and unaltered since time of acquisition (i.e. Integrity)?	◎	◎	◎	◎
	Can you validate that the origin of the content is legitimate (i.e. Authenticity)?	◎	◎	◎	◎
	Can you obtain technical documentation that establishes the origins of the content, its history to include any alterations, and chain of custody (i.e. Provenance)?	◎	◎	◎	◎
4.	DO YOU HAVE THE CAPABILITY TO ACT APPROPRIATELY WHEN PROBLEMATIC DEEPFAKES ARE DETECTED?	◎	◎	◎	◎
	Can you identify and block inappropriate content before it is ever posted?	◎	◎	◎	◎
	Can you remove content after posting?	◎	◎	◎	◎
	Can you mark, label, or flag content as potentially misleading, if you allow it to remain on your site?	◎	◎	◎	◎
	Can you take action against accounts, account owners, and content creators, including deletion or suspension, warnings, de-monitization, reduction of influence, and criminal or civil liability?	◎	◎	◎	◎
5.	ARE YOU PROACTIVELY ENGINEERING YOUR ENVIRONMENT TO REDUCE THE EFFECTS OF DEEPFAKES?	◎	◎	◎	◎
	Are you raising up authoritative voices?	◎	◎	◎	◎
	Are you rewarding trusted creators and users?	◎	◎	◎	◎
	Are you educating the public (i.e. users) on disinformation and increasing their media literacy skills?	◎	◎	◎	◎
	Are you educating your workforce?	◎	◎	◎	◎
	Are you building collaborative relationships with users, partner agencies, academia, and the public sector?	◎	◎	◎	◎

Appendix C. Examples of Current Actions Taken against Deepfakes

Establish Policies and Support Legislation

- Meta has an established policy against manipulated media and removes misleading content if 1) it has been edited – beyond adjustments for clarity or quality – in ways that aren't apparent to an average person and would likely mislead someone into thinking that a subject of the video said words that they did not actually say and 2) it is the product of artificial intelligence or machine learning that merges, replaces or superimposes content onto a video, making it appear to be authentic.[22]

- Certain types of misleading or deceptive content with serious risk of egregious harm are not allowed on YouTube, according to the company's misinformation policies. This includes certain types of misinformation that can cause real-world harm, such as media that promotes harmful remedies or treatments, incites civil unrest, or interferes with democratic processes.[23]

- Twitter has a comprehensive policy against manipulated media stating that "you may not share synthetic, manipulated, or out-of-context media that may deceive or confuse people and lead to harm ("misleading media"). In addition, we may label Tweets containing misleading media to help people understand their authenticity and to provide additional context."[24]

- Some US states created laws for certain deepfakes such as political protection during elections in California and Texas. California's law prohibits the creation of videos, images, audio of politicians doctored to resemble real footage within 60 days of an election.[25]

- In 2021 US National Defense Authorization Act (NDAA) directs Department of Homeland Security to compile and publish an annual report on the harms caused by deepfakes.[26]

- In 2021, a Deepfake Task Force was created as part of the National Defense Authorization Act for Fiscal Year 2020, turning attention onto the potential for adversary nations to use deepfakes as part of the election disinformation campaigns.[27] The purpose of the task force is to investigate policy and technology to reduce the damage of deepfakes.

Identify Deepfakes

- Social media sites use technology—such as hashes (i.e., "digital fingerprints") and machine learning algorithms—to flag content for human review teams.[28,29] For example, over 87% of the 9 million videos YouTube removed in the second quarter of 2019 were first flagged by their automated systems.[30]

- Social and news media companies also rely on content and review teams to evaluate flagged videos. The Wall Street Journal has launched an internal deepfakes task force led by the Ethics & Standards and the Research & Development teams. This group, the WSJ Media Forensics Committee, is comprised of video, photo, visuals, research, platform, and news editors who have been trained in deepfake detection.[31] Meta sends videos that do not meet their policy standards for automatic removal are still eligible for review by one of our independent third-party fact-checkers, which include over 50 partners worldwide fact-checking in over 40 languages.[32] YouTube has over 10,000 people detecting, reviewing, and removing content that violates their guidelines.[33]

- Broadcast companies also partner with users and other third parties to identify inappropriate media. In Australia, South Korea, and the United States, Twitter has tested a reporting feature that will allow users to report Tweets that seem misleading. As part of the experiment, the phrase "It's misleading" will appear as an option when you select Report an issue.[34]

- Research involving video and image deepfakes detection has been extensive. Today, research surrounding deepfake audio is also getting the attention due to the increased threat. Audio deepfakes, technically known as logical-access voice spoofing techniques, have become an increased threat on voice interfaces due to the recent breakthroughs in speech synthesis and voice conversation technologies.[35]

- The rise of AI-enabled synthetic content has been met with a commensurate increase in academic literature detailing generation and detection capabilities. specifically aimed at assessing frameworks and techniques associated with synthetic content and deepfakes. Modern deepfake detection algorithms rely on machine learning, deep learning, and statistical model analysis to evaluate content.[36] Techniques which utilize convolutional neural networks (CNNs)[e] are particularly useful for assessing content due to their ability to effectively identify, extract, and evaluate a variety of salient features of the image, including textures, facial movement, and visual artifacts. Many detection algorithms typically train on datasets composed of actual manipulated content that it is meant to detect, including content generated by the most popular and capable AI systems. Many detection algorithms achieve varying levels of success at identifying content that has been manipulated in some way.

Demonstrate Integrity, Authenticity, and Provenance

- Witness attestation or testimony is the simplest way to assert a claim of legitimacy and truthfulness of something. Relying on someone's word has had mixed results over the course of civilization, where society must depend on the integrity of a human. In modern legal proceedings, this is often insufficient to authoritatively establish the truth without corroborating support from a more objective source.

[e] CNNs are typically a neural network that analyzes visual images through processing data to detect and classify an image. An image is input in the CNN where it then analyzes the image through several algorithmic layers and data processing filters to distill into salient data and ultimately outputs a classification of the image.

- Chain of custody[f] is one of the classic examples of demonstrating the authenticity and integrity of something, such as evidentiary information.[37] This technique is a tried-and-true method regularly practiced with physical objects, including physical media containing digital content. Cyberforensics and chain of custody for digital content and data is also common, particularly with financial and health care data that is often confidential, or imaging hard drives for investigations. Yet common methods in verifying integrity and authenticity depend on exploitable datasets, such as metadata. Cryptographic hashes may likewise be insufficient to guarantee integrity of data. If contents of evidence are modified and hashes are recalculated and stored, then the modifications to the information may be undetectable, undermining efforts to establish and prove data integrity.[38]

- File hashing compares files using hash values[g] which are a generated output from an input file, such as a digital image or video file, which then could be validated through a database to verify integrity and accuracy. Several different file hashing algorithms exist for images each with their own strengths and weaknesses.[39] Some image file hashing algorithms can tolerate limited changes to an image such as compression, distortion, and inversion.[40] These algorithmic frameworks, however, have vulnerabilities to adversarial attacks as it can be easy to alter the image in a trivial, non-visual way thereby producing a different, unique[h] hash output.

Act Appropriately

- Social media companies can identify and block inappropriate content before it is ever posted. YouTube, for example, uses the hashes mentioned in the previous section to "catch copies of known violative content before they are ever made available to view."[41] Facebook is providing their customers with tools to detect and block inappropriate content.[42] It is conceivable that someday, educated consumers of multimedia content may come to expect that tools and/or proactive verification of content will be provided by the platforms they visit.

- Companies can also remove content from their site after a user has posted it. If content violates YouTube policies, they will remove it and send the channel owner an email with a warning for the first offense and a strike for additional offenses. [43] Twitter also deletes tweets for "high-severity violations of the [synthetic and manipulated media] policy, including misleading media that have a serious risk of harm to individuals or communities."[44] Meta will remove misleading manipulated media if it has been "edited or synthesized—beyond adjustments for clarity or quality—in ways that aren't apparent to an average person and would likely mislead

[f] NIST defines this as a process that tracks the movement of evidence through its collection, safeguarding, and analysis lifecycle by documenting each person who handled the evidence, the date/time it was collected or transferred, and the purpose for the transfer.

[g] A hash or hash value is simply enciphered text, which is an output resulting from some sort of data input, such as an image or text, and processed through cryptographic algorithms.

[h] Although mathematically highly improbable, "collisions" do occur, where two identical hash values exist for different input data.

someone into thinking that a subject of the video said words that they did not say" and "it is a product of AI/ML that merges, replaces, or superimposes content onto a video, making it appear to be authentic."[45]

- In some cases, companies will decide to mark, label, or flag content as potentially misleading, but allow it to remain on the site. Twitter, for example, will take a combination of actions on a tweet that violates their policy, but does not rise to the level of deletion. These actions include: applying a label and/or warning message to the Tweet; showing a warning to people before they share or like a Tweet; reduce the visibility of the Tweet and/or prevent it from being recommended; turning off likes, replies, and Retweets; and/or providing a link to additional explanations or clarifications. [46] Meta reserves the right to fact-check media that does not meet the standards for removal. If their independent fact-checkers rate the media as false or partly false, Meta will significantly reduce its distribution in News Feed, reject it if it is being run in an ad, or post warnings that the media is false to people who try to or have already shared it. They say that leaving a video up and labeling it as false will provide people with important information and context, especially since the video might still be available elsewhere on the Internet.[47]

- Companies have also come up with a variety of actions against accounts and account owners, including deletion or suspension, warnings, de-monitization, and reduction of influence. YouTube gives channels a warning and three strikes for instances of content that violate policies before terminating the channel.[48] Twitter will temporarily reduce the visibility of, lock, or suspend an account that has advanced or continuously shared misleading narratives that violate their policies.[49]

Engineer the Environment

- Raise up authoritative voices and reward trusted creators and users. YouTube, for example, can adjust its algorithm to raise authoritative sources, such as trusted news sites and recognized experts, to the top of their recommendations.[50] Also, in order to monetize, "channels must also comply with the YouTube channel monetization policies, which includes our Advertiser-friendly content guidelines which do not allow ads on content promoting or advocating for harmful health or medical claims; or content advocating for groups which promote harmful misinformation. Violation of our YouTube channel monetization policies may result in monetization being suspended," according to YouTube.[51] Twitter also sources and elevates relevant context from reliable sources. [52,53]

- Broadcast companies have attempted to educate the public on disinformation and increase media literacy skills. YouTube helps users build media literacy skills; enables the work of organizations who work on media literacy initiatives; and invests in thought leadership to understand the broader context of misinformation.[54]

- Education of the media workforce can also help catch deepfakes before they spread. The Wall Street Journal (WSJ), for example, has launched an internal deepfakes task force led by the Ethics & Standards and the Research & Development teams. This

group, the WSJ Media Forensics Committee, is "comprised of video, photo, visuals, research, platform, and news editors who have been trained in deepfake detection. Beyond this core effort, they are hosting training seminars with reporters, developing newsroom guides, and collaborating with academic institutions such as Cornell Tech to identify ways technology can be used to combat this problem."[55]

- Building collaborative relationships with users, partner agencies, academia, and the public sector can inform policy development, improve technology, and increase awareness. Meta has been driving global conversations with technology, policy, media, legal, civic and academic experts to inform their actions. They have also partnered with Reuters, the world's largest multimedia news provider, to "help newsrooms worldwide to identify deepfakes and manipulated media through a free online training course. News organizations increasingly rely on third parties for large volumes of images and video, and identifying manipulated visuals is a significant challenge."[56] Twitter has developed close partnerships with AP, Reuters, and other news agencies; public health authorities; and governments to consult and get various media and claims reviewed.[57]

    - Broadcast companies can partner with voices on an international platform such as the United Nations, to prioritize advocacy on the prevention of deepfakes in the context of 'technology facilitated gender-based violence.'[58] This advocacy against deepfakes is progressing towards a demand for copyright privileges on body- images from a human rights perspective, thereby increasing the seriousness of addressing deepfakes. The United Nations Population Fund (UNFPA) has modeled a strategy to highlight technology facilitated gender-based violence by partnering with Sensity AI in providing an estimate on the total number of deepfakes generated every second.[59]

    - Large media companies could partner with academia and government to encourage the tackling of deepfakes through media literacy, research, and school projects. Academia has a dual impact in both empowering young-minds about the ill-effects of deepfakes in addition to sculpting a pathway for their participation in developing solutions. For example, the state of Illinois amended their school code by adding a provision that, beginning with the 2022-2023 school year, every public high-school is required to include in its curriculum a unit of instruction on media literacy.[60] Google—through its News Literacy Project—and Twitter—through a partnership with the United Nations Educational, Scientific, and Cultural Organization–provide media literacy education to students, educators, and the public.[61,62]

Appendix D. Current Efforts in ICAM Technologies

ICAM systems are generally concerned with two forms of access: physical and logical. The former refers to access of physically defined areas or spaces, such as the sterile areas at an airport beyond the checkpoint. Logical access refers to all logically defined access methods including enterprise access to employee accounts and information, consumer web or mobile login to financial services organization accounts, and the like. Access privileges are granted after verifying the identity assertion made by the person using the system or service. Once the assertion is verified, the person is allowed access to one or more services at authorization levels determined by rules managed within the ICAM system.

Most modern ICAM technologies use MFA to reduce risk, and this risk certainly extends to the threats from deepfakes. When using multiple authentication factors to verify an identity claim, the combined strength of the factors, each of which alone may be less than 100% perfect and accurate, would be infinitesimally smaller than each of the factor alone. Experts from the Department of Homeland Security studying identity risk have noted that a combination of factors used in this way to establish one's identity reduce that risk and the assurance coming from a strong identity extends to the verification process too. Now let us examine some authentication factors and attack vectors from the perspective of deepfakes.

- ***Shared Secret:*** While a category of cryptography uses a shared secret for secure sharing of data between different entities (this is also known as symmetric encryption, as opposed to asymmetric encryption, or public key cryptography), perhaps the most commonly encountered form of a shared secret is a password, or a keyword known only to the identity claimant. These can be easily compromised, guessed, or difficult to remember across all the platforms and services to which the user needs access. Implementing policies to change passwords periodically, and increasingly frequently based on the level security needed, the risk of compromise can be reduced. In a video or audio environment for verification, these passwords or one-time digital tokens can be requested using the user interface of the application, or simply consumed as utterances that are verified using speech recognition. Note that speech recognition is different from speaker recognition and refers only to the transformation of audio signals into text.

- ***Physical Token:*** In addition to passwords or one-time passcodes, physical tokens, while burdensome, can also be used to compound the strength of authentication as an additional you-have factor. Physical tokens can of course be stolen or misplaced, and unless they themselves are further protected by a password or a biometric, they are vulnerable for exploitation in the wrong hands. Physical tokens are otherwise a reasonable defense against deepfakes, introducing a human-in-the-loop physical element to the verification process.

- ***Identity Document:*** We have learned that experts are focusing on identity documents as a potential vector for fraudulent identity claims. While the only possibility of

attacking this factor seems to be via the insertion of a deepfake image into the document, experts believe that validating other elements of the identity document and all the identity attributes presented in the document itself, combined, offer a strong defense against fraudulent identity proofing and issuance.

- *Token-based Address Verification:* While likely not a very practical scenario in common and frequently-performed identity verification transactions, token-based verification in one's home address, driven by, as an example, the scanning of a QR code on a mobile application which can geo-fence the transaction for security and verify that the individual receiving the code is doing so at their address of residence, can offer a high level of assurance for identity proofing, registration, or infrequent verification processes.

- *Biometrics:* Biometrics is generally considered and utilized as a strong you-are factor of authentication. As biometrics, especially face biometrics, rapidly approaches commodity status with a broad range of authentication applications using biometrics as an additional, or sometimes the primary, factor of authentication, vulnerabilities to fraud have emerged. With an increasing volume of remote transactions in unmonitored environments as we previously discussed, presentation attacks are amongst the most commonly expected threats and many solution developers are actively working to develop presentation attack or liveness detection components to increase the strength of their authentication systems. The National Institute of Standards and Technology has recently announced that they will offer tests to evaluate facial presentation attack detection algorithms under a new benchmark track called FRVT PAD (Face Recognition Vendor Technology Presentation Attack Detection), which will provide ongoing independent testing of software-based facial PAD detection technologies[63], closing an important gap in standardized testing and evaluation of PAD-related technology. Images or audio generated via deepfakes can also be part of a presentation attack, although currently the industry is moving more slowly against this particular threat.

- *Challenge Response:* A well-known method to defend actively against presentation attacks (as opposed to passive defenses which rely on automated detection technologies), challenge responses that require the active engagement of the identity claimant are useful, and could certainly stop attacks such as spoofing attacks or deepfakes even if they are successfully presented real-time. Specific random gestures, mimics or other artifacts requested from the user can both be difficult to create real-time in a video or audio sample, and even when it is, the generation technology typically leaves traces that can be detected automatically, exposing the liveness or deepfake attack.

  A variant of the challenge response approach which has application to video calls involves the use of real-time manipulation of one's environment to evoke a

detectable response within the environment of the individual being challenged. In the typical video call, users are facing both a camera and a screen displaying the individual(s) with whom they are meeting. Should the browser window depicting one individual suddenly emit a bright flash of light, the viewer's face would be expected to reflect that flash of light with a visible change in brightness. (See Figure 4).

Figure 4. A change in the illumination level of a browser window can change the visual appearance of a participant in a video conference call. This figure shows the difference that results from changing a browser window from dark (left) to bright (right).[64]

If, however, the viewer was projecting a deepfake, then the deepfake face would not display a change in brightness. This technique is referred to as "Active Illumination.[65]" Randomly varying the color of the light flashes in this scenario could make this technique even harder to defeat.[66]

Appendix E. Education & Awareness Efforts to Mitigate Deepfakes

The rapid emergence of deepfakes and other synthetic media over the last several years has given the phrase "seeing is believing" a limited shelf life. It has become apparent to many in the government and private sectors that individuals need to be better prepared to deal with the potential use and abuse of synthetic media. As a result, multiple efforts have emerged to better prepare decision makers – whether a member of the public or an individual in an organization providing a service.

- To prepare individuals to address concerns about synthetic media, they must first be aware of the issue. Educational programs can provide individuals with basic understanding to the problem and can start at an early age. Among the many programs that speak to education of children are curricula guides for children from kindergarten to high school developed by Common Sense, a non-profit. https://www.commonsense.org/education/videos/deepfakes-and-democracy

- Local agencies such as the Calvert County (Maryland) Public Schools also offered programming curricula related to deepfakes: https://programminglibrarian.org/blog/deepfakes-part-2-resources-all-ages.

- Similarly, the Massachusetts Institute of Technology (MIT) Media Laboratory offers online content in media literacy that can be incorporated into formal curricula or accessed independently: https://news.mit.edu/2022/fostering-media-literacy-age-deepfakes-0217.

While the above efforts focus on awareness within the context of primary and secondary education, a critical need also exists to educate people beyond high school.

- For example, one community considered to be highly vulnerable to online deception is senior citizens. To address the needs of this community the Center for Information Integrity at the University at Buffalo created the Deception Awareness and Resilience Training (DART). Using funding from the National Science Foundation Convergence Accelerator program, DART "… is a cross-disciplinary, multi-institutional, user-centered effort to develop engaging, effective, gamified tools to build resilience to disinformation in older adults." https://www.buffalo.edu/cii/projects/DART.html

The DART program address deepfakes within the broader context of disinformation awareness, but for those interested specifically in deepfake image and video detection, there are multiple online resources, such as:

- The SANS Institute's March 1, 2022 newsletter entitled "Learn a New Survival Skill: Spotting Deepfakes" (available at https://www.sans.org/newsletters/ouch/learn-a-new-survival-skill-spotting-deepfakes/).

- Likewise, the MIT Media Lab mentioned above offers a website entitled “Detect DeepFakes: How to counteract misinformation created by AI” webpage (available at https://www.media.mit.edu/projects/detect-fakes/overview/).

Many individuals and organizations interested in fighting misinformation and disinformation are creating websites and tools that help users verify content they encounter online.

- The RAND Corporation has created an online database that identifies such resources (Available at:https://www. rand. org/research/projects/truth-decay/fighting-disinformation/search. html) Many of these tools, such as “Captain Fact” and “ClaimBuster,” are designed as collaborative environments where users work together to fact check internet content. One such collaborative site, “Snopes.com” has been in operation since 1995, when it started as “The Urban Legends ReferencePages.”

- While the value of crowd sourcing for content authentication cannot be underestimated, many individuals may just seek a tool that can help them better understand a specific piece of content, such as a photo or video. The RAND database includes some tools that provide a narrowly-focused capability, such as“Get-Metadata Viewer” and “Forensically Image Verification Tool,” which are specifically addressing the veracity of photographs.

Such targeted tools are constantly being developed in the research community and many of them apply a variety of image analysis techniques to detect inconsistencies in the biological, biometric, or physiological content in videos of people. A high level overview of how some of these techniques work is included in a recent Wired article “AI-Based Tools for Detecting Deepfakes” (Available at:https://industrywired. com/ai-based-tools-for-detecting-deepfakes/).

End Notes

[1] (U) | Gabriella Swerling | The Telegraph | https://www.telegraph.co.uk/news/2020/01/31/deepfake-audio-used-custody-battle-lawyer-reveals-doctored-evidence/ | 31 January 2020 | "Doctored audio evidence used to damn father in custody battle"
[2] (U) | Rachel Metz | CNN | https://www.cnn.com/2022/03/16/tech/deepfake-zelensky-facebook-meta/index.html | | 16 March 2022 | "Facebook and YouTube say they removed Zelensky deepfake,"
[3] (U) | Inside Edition | https://www.youtube.com/watch?v=enr78tJkTLE | 17 March 2022 | "Deepfake of Zelenskyy Tells Ukrainian Troops to Surrender"
[4] (U) | Jackson Cote | News @ Northeastern | https://news.northeastern.edu/2022/04/01/deepfakes-fake-news-threat-democracy/ |1 April 2022 | "Deepfakes and fake news pose a growing threat to democracy, experts warn."
[5] (U) | The Institute of Politics | University of Chicago | https://disinfo2022.com/ | "Disinformation and the Erosion of Democracy; In What Happens When We Can't Tell What's Real?"
[6] (U) | A Anand & Bianco, B. | United Nations Institute for Disarmament Research | 2021 | "The 2021 Innovations Dialogue Conference Report: Deepfakes, Trust and International Security, Geneva, Switzerland: United Nations Institute for Disarmament Research [UNIDIR]".
[7] (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 | 8 July 2020 | "Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios"
[8] (U) |US General Services Administration | https://www.gsa.gov/policy-regulations/policy/information-integrity-and-access/identity-credential-and-access-management#:~:text=Identity%2C%20credential%2C%20and%20access%20management%20(ICAM)%20comprises%20the,secure%20access%20to%20protected%20resources | "Identity Assurance and Trusted Access"
[9] (U) | National Institute of Standards and Technology | https://pages.nist.gov/SOFA/SOFA.html | "Strength of Function for Authenticators – Biometrics (SOFA-B)"
[10] (U) | ISO | https://www.iso.org/standard/67381.html | 2017 | | "Information technology – Biometric presentation attack detection – Part 3: Testing and reporting"
[11] (U) | Morris Dworkin | National Institute for Standards and Technology | https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-38b.pdf | May 2005 | "Recommendation for Block Cipher Modes of Operation"
[12] (U) | Jon Boyens, et al | National Institute of Standards and Technology | https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-161r1.pdf | May 2022 | "Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations"
[13] (U) | Scott Briscoe | ASIS International | https://www.asisonline.org/security-management-magazine/latest-news/today-in-security/2021/january/U-S-Laws-Address-Deepfakes/ | 12 January 2021 | "U.S. Laws Address Deepfakes"
[14] (U) | Alyssa Ivancevich | Hastings Constitutional Law Quarterly | https://repository.uchastings.edu/cgi/viewcontent.cgi?article=2150&context=hastings_constitutional_law_quaterly | February 2022 | "Deepfake Reckoning: Adapting Modern First Amendment Doctrine to Protect Against the Threat Posed to Democracy"
[15] (U) | Alyssa Ivancevich | Hastings Constitutional Law Quarterly | https://repository.uchastings.edu/cgi/viewcontent.cgi?article=2150&context=hastings_constitutional_law_quaterly | February 2022 | "Deepfake Reckoning: Adapting Modern First Amendment Doctrine to Protect Against the Threat Posed to Democracy"
[16] (U) | Meta |https://ai.facebook.com/datasets/dfdc/ | 25 June 2022 | "Deepfake Detection Challenge Dataset"
[17] (U) | Paravision | https://www.paravision.ai/news/why-we-are-tackling-deepfakes/ | June 2022 | "Why we are tacking deepfakes,"
[18] (U) | Rubel Biswas and Pablo Blanco-Medina | Arxiv | https://arxiv.org/pdf/2108.11794.pdf. | 26 Aug 2021| "State of the Art: Image Hashing"

[19] (U) | Cameron C. White and Manoranjan Paul | Arxiv |https://arxiv.org/ftp/arxiv/papers/2004/2004.06860.pdf | 5 April 2020 | "A Practical Blockchain Framework using Image Hashing for Image Authentication"
[20] (U) | Saltz et al. | Medium.com | https://medium.com/swlh/it-matters-how-platforms-label-manipulated-media-here-are-12-principles-designers-should-follow-438b76546078 | 9 June 2020 | "It Matters How Platforms Label Manipulated Media. Here are 12 Principles Designers Should Follow"
[21] (U) | Regional Computer Forensics Laboratory | https://www.rcfl.gov/about
[22] (U) | Meta | https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/ | 6 January 2020 | "Enforcing Against Manipulated Media"
[23] (U) | YouTube | https://support.google.com/youtube/answer/10834785 | "Misinformation Policies"
[24] (U) | Twitter | https://help.twitter.com/en/rules-and-policies/manipulated-media | "Synthetic and Manipulated Media Policy"
[25] (U) | Scott Briscoe | ASIS International | https://www.asisonline.org/security-management-magazine/latest-news/today-in-security/2021/january/U-S-Laws-Address-Deepfakes/ | 12 January 2021 | "U.S. Laws Address Deepfakes"
[26] (U) | Alyssa Ivancevich | Hastings Constitutional Law Quarterly | https://repository.uchastings.edu/cgi/viewcontent.cgi?article=2150&context=hastings_constitutional_law_quaterly | February 2022 | "Deepfake Reckoning: Adapting Modern First Amendment Doctrine to Protect Against the Threat Posed to Democracy"
[27] (U) | Jule Pattison-Gordon | Government Technology | https://www.govtech.com/security/senate-committee-advances-bill-to-create-deepfake-task-force | 6 August 2021 | Senate Committee Advances Bill to Create Deepfake Task Force
[28] (U) | Twitter | https://help.twitter.com/en/rules-and-policies/manipulated-media | "Synthetic and Manipulated Media Policy"
[29] (U) | YouTube Official Blog | https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove/ | 3 September 2019 | "The Four Rs of Responsibility, Part 1: Removing Harmful Content"
[30] (U) | YouTube Official Blog | https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove/ | 3 September 2019 | "The Four Rs of Responsibility, Part 1: Removing Harmful Content"
[31] (U) | Francesco Marconi and Till Daldrup | Nieman Lab | https://www.niemanlab.org/2018/11/how-the-wall-street-journal-is-preparing-its-journalists-to-detect-deepfakes/ |15 November 2018 | "How The Wall Street Journal is preparing its journalists to detect deepfakes"
[32] (U) | Meta | https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/ | 6 January 2020 | "Enforcing Against Manipulated Media"
[33] (U) | YouTube Official Blog | https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove/ | 3 September 2019 | "The Four Rs of Responsibility, Part 1: Removing Harmful Content"
[34] (U) | Twitter | https://help.twitter.com/en/rules-and-policies/manipulated-media | "Synthetic and Manipulated Media Policy"
[35] (U) | Tianxiang Chen, et al | Odyssey 2020 The Speaker and Language Recognition Workshop | https://www.isca-speech.org/archive_v0/Odyssey_2020/pdfs/29.pdf | 1 November 2020 | "Generalization of Audio Deepfake Detection" | 2020.
[36] (U) | Shohel Rana, et al | Digital Object Identifier | https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9721302 |10 March 2022 | "Deepfake Detection: A Systematic Literature Review"
[37] (U) | Wayne Jansen and Richard Ayers | National Institute for Standards and Technology | https://csrc.nist.gov/publications/detail/sp/800-72/final | November 2004 | "Guidelines in PDA Forensics"
[38] (U) | Makhdoom Syed Muhammad Baqir Shah, et al | The Journal of Digital Forensics, Security and Law | https://commons.erau.edu/cgi/viewcontent.cgi?article=1478&context=jdfsl | 30 June 2017 | "Protecting Digital Evidence Integrity and Preserving Chain of Custody"
[39] (U) | Cameron C. White and Manoranjan Paul | Arxiv | https://arxiv.org/ftp/arxiv/papers/2004/2004.06860.pdf |5 April 2020 | "A Practical Blockchain Framework using Image Hashing for Image Authentication"
[40] (U) | R Venkatesan, et al | Proceedings 2000 International Conference on Image Processing | https://ieeexplore.ieee.org/document/899541/ |2000 | "Robust Image Hashing"
[41] (U) | YouTube Official Blog | https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove/ | 3 September 2019 | "The Four Rs of Responsibility, Part 1: Removing Harmful Content"

[42] (U) | TechXplore | https://techxplore.com/news/2022-03-facebook-tools-misinformation-users-groups.html | 9 March 2022 | "New Facebook tools target misinformation in users' groups"
[43] (U) | YouTube | https://support.google.com/youtube/answer/10834785 | "Misinformation Policies"
[44] (U) | Twitter | https://help.twitter.com/en/rules-and-policies/manipulated-media | "Synthetic and Manipulated Media Policy"
[45] (U) | Meta | https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/ | 6 January 2020 | "Enforcing Against Manipulated Media"
[46] (U) | Twitter | https://help.twitter.com/en/rules-and-policies/manipulated-media | "Synthetic and Manipulated Media Policy"
[47] (U) | Meta | https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/ | 6 January 2020 | "Enforcing Against Manipulated Media"
[48] (U) | YouTube | https://support.google.com/youtube/answer/10834785 | "Misinformation Policies"
[49] (U) | Twitter | https://help.twitter.com/en/rules-and-policies/manipulated-media | "Synthetic and Manipulated Media Policy"
[50] (U) | YouTube Official Blog | https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove/ | 3 September 2019 | "The Four Rs of Responsibility, Part 1: Removing Harmful Content"
[51] (U) | YouTube | https://www.youtube.com/howyoutubeworks/our-commitments/fighting-misinformation/#rewarding-trusted-creators-and-artists | "How does YouTube address misinformation? Rewarding Trusted Creators and Artists."
[52] (U) | Twitter | https://help.twitter.com/en/rules-and-policies/manipulated-media | "Synthetic and Manipulated Media Policy"
[53] (U) | Twitter Official Blog| https://blog.twitter.com/en_us/topics/company/2021/bringing-more-reliable-context-to-conversations-on-twitter | 2 August 2021 | "Bringing More Reliable Context to Conversations on Twitter"
[54] (U) | YouTube | https://www.youtube.com/howyoutubeworks/our-commitments/fighting-misinformation/#media-literacy | "How does YouTube address misinformation? Media Literacy."
[55] (U) | Francesco Marconi and Till Daldrup | Nieman Lab | https://www.niemanlab.org/2018/11/how-the-wall-street-journal-is-preparing-its-journalists-to-detect-deepfakes/ |15 November 2018 | "How The Wall Street Journal is preparing its journalists to detect deepfakes"
[56] (U) | Meta | https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/ | 6 January 2020 | "Enforcing Against Manipulated Media"
[57] (U) | Twitter Official Blog| https://blog.twitter.com/en_us/topics/company/2021/bringing-more-reliable-context-to-conversations-on-twitter | 2 August 2021 | "Bringing More Reliable Context to Conversations on Twitter"
[58] (U) | UNFPA | https://www.unfpa.org/publications/technology-facilitated-gender-based-violence-making-all-spaces-safe | 2021 | "Technology-facilitated Gender-based Violence: Making All Spaces Safe"
[59] (U) | UNFPA | https://www.unfpa.org/bodyright | "Introducing a new copyright for the human body"
[60] (U) | Illinois Civics Hub | https://www.illinoiscivics.org/standards/media-literacy-toolkit/#:~:text=IL%20House%20Bill%20234%20amended,of%20instruction%20on%20media%20literacy | "Media Literacy Toolkit"
[61] (U) | Google Official Blog | https://blog.google/outreach-initiatives/google-news-initiative/media-literacy-partnerships/ |16 November 2021 | "Supporting media literacy with new partnerships"
[62] (U) | Twitter Official Blog | https://blog.twitter.com/en_us/topics/company/2019/twitter-launches-new-media-literacy-handbook-for-schools |24 October 2019 | "Twitter builds partnership with UNESCO on media and information literacy"
[63] (U) | National Institute for Standards and Technology | https://pages.nist.gov/frvt/html/frvt_pad.html | "FRVD PAD"
[64] (U) Candice R. Gerstner and Hany Farid | Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops | pp. 53-60 | 2022 | "Detecting Real-Time Deep-Fake Videos Using Active Illumination"
[65] (U) Ibid.
[66] (U) Ibid.