Guide for Preparing and Responding to Deepfake Events

From the OWASP Top 10 for LLM Applications Team

Version: 1

Published: **September 2024**

Revision History

Revision	Date	Authors	Description
.01	June 28, 2024	Rachel James Bryan Nakayama	First Draft
.03	July 9th, 2024	Rachel James Bryan Nakayama Sarah Thorton Ramesh Kumar Vaibhav Malik	Feedback and Second Draft
.05	August 6th, 2024	Rachel James Bryan Nakayama Sarah Thorton Ramesh Kumar Vaibhav Malik Manuel Villanueva	Fourth Draft
1	September 10, 2024	Rachel James Bryan Nakayama	Published

License and Usage

Contents

Note from OWASP CTI Layer Lead Authors

Early in 2024, the OWASP Top 10 for LLM & GenAI community expressed a great deal of interest in covering the adversarial use of artificial intelligence. On reflection, the core team determined that since vulnerabilities within AI systems was the primary focus of the Top 10, guidance on adversarial use fell outside the scope of that publication. The OWASP community came together and volunteered to create a separate resource group, one that would focus on creating actionable guidance, checklists and research into adversarial use for cybersecurity professionals. That group became known as the CTI Layer team, led by Rachel James and Bryan Nakayama.

This Guide for Preparing and Responding to Deepfake Events is the first of several planned publications of the CTI Layer. This publication was developed by cybersecurity professionals for cybersecurity professionals. We intended to provide practical guidance for a technical and leadership audience who must create playbooks, response plans and quickly respond to a deepfake event. It is the hope of the authors and contributors of this document that we can improve the cultivation of best practices and provide a comprehensive overview of what is involved in preparing, detecting, and responding to such events.

Rachel James, CISSP, CISA, OSCP, GMLE

CyberShujin LLC
racheljames@cybershujin.com

Bryan Nakayama, Ph.D.

CTI Professional
bryan.nakayama@owasp.org

Overview

Deepfakes—hyper-realistic digital forgeries—have gained significant attention as the rapid development of generative AI has made it easier to produce convincingly realistic videos and audio recordings that can deceive even the most discerning viewers. They pose a potentially daunting challenge for cybersecurity professionals since fraudsters and cybercriminals can leverage deepfakes to carry out sophisticated impersonation and social engineering attacks. Due to the widespread use of social media, everyone from high-profile individuals like CEOs to average citizens are at the risk of impersonation since it can take as little as 10 seconds of audio or video to produce a convincing deepfake. Deepfake-generated content has already been used in phishing and fraud schemes, where attackers created videos of CEOs and other trusted figures to manipulate employees into divulging sensitive information and/or transferring funds (Chen & Magramo, 2024).

While deepfakes are a powerful tool for social engineering, cybersecurity professionals do not need to turn to new detection technologies or intensive "how to spot a deepfake" training programs in order to mitigate the risk that they pose. Recent studies suggest that deepfake detection technologies are still immature and the rapid advance of the technology will make training programmes focused on looking for specific visual or audio artifacts rapidly out-of-date (GAO, 2024). Moreover, researchers have discovered that even with training people both cannot reliably detect deepfakes and tend to overestimate their own ability to identify deepfakes (Köbis et al., 2021). Like many other social engineering attacks, deepfake-enhanced attacks frequently depend on the victim bypassing established procedures and controls at the behest of the attacker. Therefore, this guide emphasizes practical and pragmatic defense-in-depth strategies as well as layered controls as a key approach that cybersecurity professionals should take to deepfakes.

The hope is to provide a guide that is resilient to evolving deepfake-enhanced threats by applying fundamental security principles. Key strategies that the guide endorses include:

- Focusing on process adherence rather than visual or auditory detection of fakes.
- Implementing and maintaining strong financial controls and verification procedures.
- Cultivating a culture of awareness and skepticism towards unusual requests.
- Developing and regularly updating incident response plans.

The first section in the guide is the Scope which outlines key definitions and the intended audience. The guide distinguishes between four different scenarios based on attacker intentions — financial fraud, job interview fraud, social engineering, mis/dis/malinformation — and provides guidance across the four stages of incident response from NIST Special Publication 800-61 Revision 3:

1. Preparation,
2. Detection and Analysis,
3. Containment, Eradication and Recovery
4. Post-incident activity

Scope

Synthetic media intended to reproduce someone's likeness is generally divided into two categories (DOD, 2023):

- **Cheapfakes** - Multimedia that has been manipulated using techniques that do not involve machine/deep learning, which in many cases can still be as effective as the more technically sophisticated techniques, are often referred to as shallow or cheap fakes.
- **Deepfakes** - Multimedia that have either been created (fully synthetic) or edited (partially synthetic) using some form of machine/deep learning (artificial intelligence) are referred to as deepfakes.

We will focus this guidance on three categories of malicious deepfakes based on the attacker's objective:

1. Financial gain through fraud by impersonation
2. Job interview fraud
3. Impersonation to further cyberattacks (such as initial access)
4. Mis/Dis/Mal information

Most organizations outside government and journalistic entities will potentially be targeted for one of these three objectives. Based on public and private intelligence sources, we believe there has been a minor but measurable increase in activity from these three categories impacting organizations since mid-2023.

We separate malicious deepfake activity into three categories because the preparation and response for each are different. For example , if a threat actor attempts to use a deepfake for fraud or trick a help desk employee into giving them access, you are unlikely to be fortunate enough to have any captured video or audio to analyze, nor will the content be hosted on a platform. In the case of mis/dis/mal information, there will likely be some media to analyze and a take-down process to be conducted.

While this guide provides preparation guidance that is encompassing of all three categories of deepfake events, the subsequent Detection and Analysis; Containment, Eradication and Recovery, and Post-Incident Activity guidance are event-specific.

Preparation

Preparation focuses on understanding the current risk through analysis of threat activity and current defensive posture, establishing a deepfake incident response plan, and finally, establishing a deepfake reporting process and employee education.

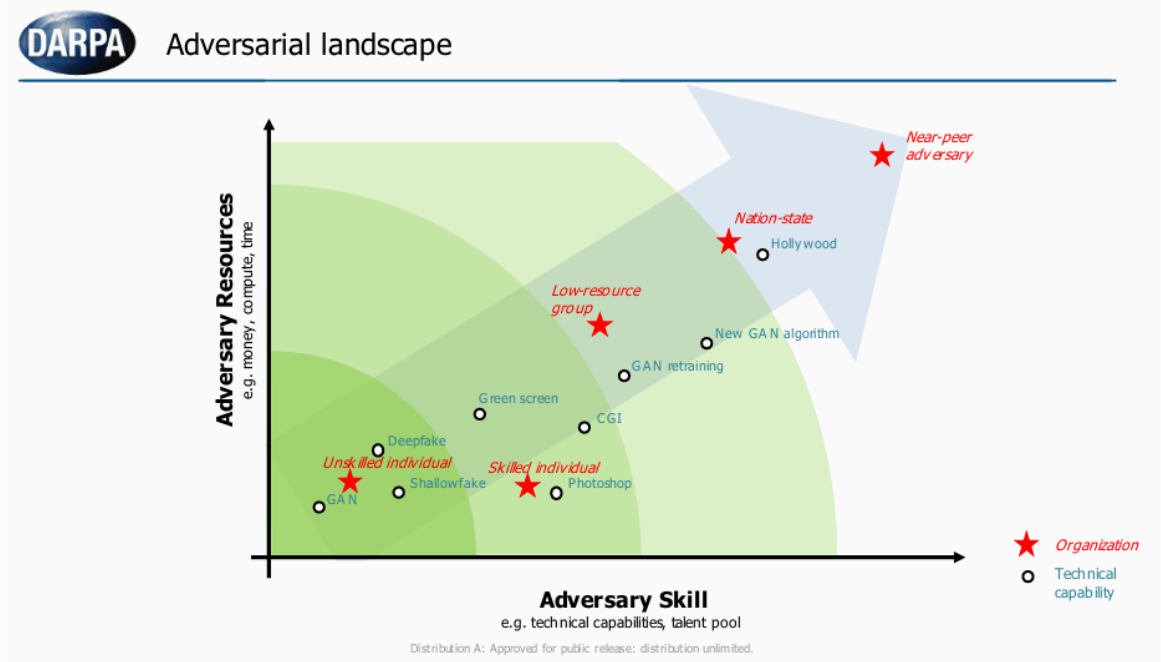
Risk assessment

At the time of writing (July 2024), deepfakes are not the leading cause of fraud, cyber threat activity or reputational damage for most organizations outside of journalistic and government entities. That said, public and private cyber threat intelligence sources indicate that the use of deepfakes and cheapfakes by financially motivated threat actors has seen a minor increase. As the technology progresses, it will become easier and cheaper to create a deepfake that is good enough for broad attacks (Ciancaglini & Sancho, 2024).

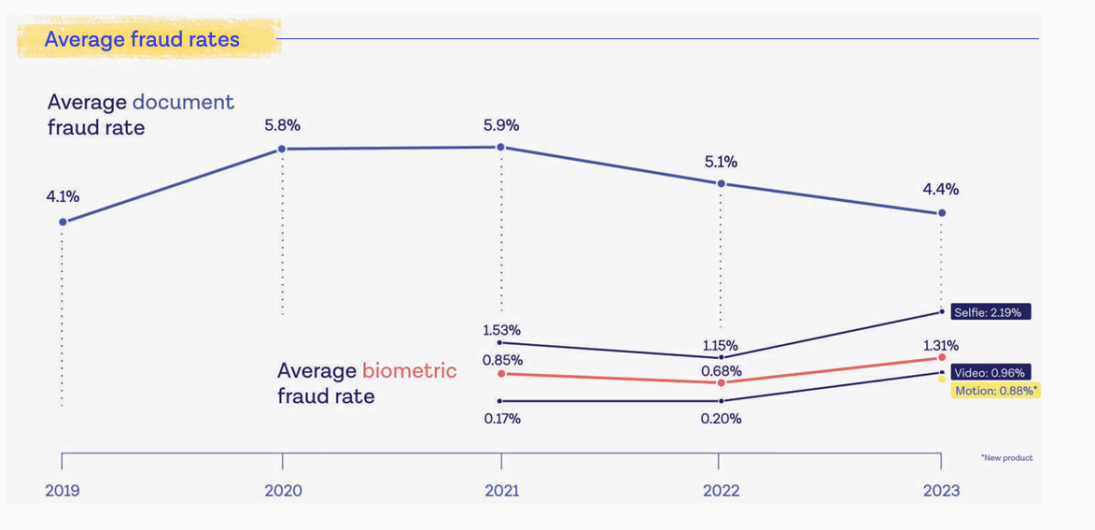
As part of preparation for deepfake events, organizations should evaluate their own individual risk of becoming a target of deepfakes based on their business, media and political exposure, history, threat actor activity, and susceptibility to fraud.

Threat Actors

An excellent adversarial landscape graphic produced by DARPA (Brooks et al., 2022), shown below, provides an understanding of the adversarial skill level and the observed use of these technologies. This will help organizations better determine which category of deepfake technical capability they will most likely encounter.



This research also reflects findings in the work produced by Onfido in collaboration with FIDO Alliance, in the *2024 Identity Fraud Report.*



The uptick in biometric fraud and use of cheapfakes and deepfakes as a means to bypass authentication and commit fraud has had a minor observable uptick. The authors conclude that this trend has seen a slight increase and that we can expect that increase to continue. Therefore, this window of opportunity for cybersecurity professionals to develop awareness, detection, response, and mitigation strategies is ideal.

Threat Activity

Current, known threats that these technologies pose include:

1. Evasion of authentication - How I Broke Into a Bank Account With an AI-Generated Voice
2. Impersonation - Unusual CEO Fraud via Deepfake Audio Steals US$243,000 From UK Company
3. Financial Fraud - Finance worker pays out $25 million after video call with deepfake 'chief financial officer'
4. Reputational damage - a fake but realistic video of a CEO making unsavory comments or incorrect statements can damage brand image and lead to loss Beware of deepfake of CEO recommending stocks, says India's National Stock Exchange and fake twitter accounts causing losses to company Eli Lilly and Lockheed Martin Responding to Malicious Corporate Deepfakes – Debevoise Data Blog
5. Deepfake employment interviews - Criminals Use Deepfake Videos to Interview for Remote Work
6. Misinformation leading to financial implications - Such as impacting stock prices: S&P Sheds $500 Billion from Fake Pentagon Explosion

Assessment of Defenses

Your assessment should include a review of policies, procedures, enforcement and auditing methods for four main areas: sensitive data disclosure, helpdesk, financial transactions and event response.

We recommend starting with a review of the governance and approval structures to oversee security measures and policies related to sensitive data disclosure, mergers and acquisitions, legal, financial transactions, and employee identification for purposes of authorization or identification (such as with the helpdesk, HR, and physical security). A key part of this review should include interviewing employees enacting these processes in order to understand whether and to what extent there are deviations from policy. Starting with this review will allow you to be able to navigate

successfully through the governance and approval structures to suggest changes and posture hardening processes.

Human-Based Authentication Best Practices

Ideally, at least two of the following best practices where human-based authentication is permitted should be in use. These best practices include:

- Maintain an employee directory of approved communication methods that can act as additional verification to authenticate a user such as corporate instant messenger, additional phone number, alternative emails or aliases that can be used to confirm a voice request.
- Alternative Communication Verification: calling the person back on a pre-registered phone number to confirm the identity and request.
- Code of the Day - in this practice, often implemented in financial institutions, requires the caller or requester to refer to a secure system which generates a random unique code that rotates on a frequent basis. Despite its name, the Code of the Day typically rotates several times a day and is used in conjunction with other forms of verbal identification. Some organizations use a secure application that requires MFA to access the current code, while others distribute the code through SMS. Users must have the ability to request a rotation of the code or to report a suspected compromise of the code, to allow for rotation-on-demand in addition to frequent automatic rotations. In situations where an employee is unable to authenticate to the application or device to get the code, it is permissible for a manager or coworker to share the code only in person and after confirming a valid employee badge (by swiping into a secure area).
- Custom Security Questions: established when onboarding, or created for third-parties and kept in encrypted storage. These should not be any data that is able to be derived from a credit report, social media account, or that the employee uses routinely (date of birth, employeeID, employee login name, should not be used). Disallow common questions such as "Mother's Maiden name" or "Pet's first name."
- Require the caller's manager or supervisor to verify the request by sending an email, or conducting an outbound call to the manager on a pre-registered phone number.

Financial Transactions

Ensure the following best practices for financial transactions are contained in written policy and procedure, and have means of enforcement and auditing for failures:

- Clear written policies regarding financial transactions and controls.
- SoD (Separation of Duties): Separate critical functions so that no single individual has control over all aspects of any financial transaction. For example, the person who authorizes a payment should differ from the person who processes it, and both should have independent non-overlapping decision-making/justification chains to do their parts respectively.
- Dual Authorization: Require two authorized individuals to approve significant transactions. This ensures that every person can only initiate and complete a transaction with oversight.
- Consider a "code of the day" technique that must be stated for any authorization of transactions or sharing of sensitive information. Accessing the day's code requires both parties to access a portal displaying the code.
- MFA on all systems for communication and financial transaction processing.
- Identify processes that permit authorization and authentication through means that are not protected by MFA.
- Inventory the method of human-based authentication, and review for best practices.
- Dual-band communication verification requires two types of authentication that cannot come through a single communication channel. For example, Transactions

should not be able to be requested, reviewed, or approved solely through email or phone calls.
- Regular audits and periodic access reviews to ensure the above.
- Ensure compliance procedures for financial transactions provide significant latitude to challenge senior leadership requests.
- Require multiple approvals for transactions above a certain threshold.

Helpdesk

- Review current policy and procedures for password reset, new device enrollment into MFA, and reporting for repeated failed verbal authentication attempts.
- Interview employees in those departments to determine current workflow (which maybe different than documented policy or process).
- Test process (after obtaining permission to do so).
- Identify gaps in policy, procedure and actual practice.
- Identify processes that permit authorization and authentication through means that are not protected by MFA.
- Inventory the method of human-based authentication, review for best practices.

Hiring

- Ensure there is an established process for reporting suspicion of candidate impersonation or fraud, and that all recruiters and hiring managers receive awareness training about the trends and the reporting process.
- Review identification verification processes for new employees. Ensure there is enhanced verification of IDs for all applicants to detect forged identities. Consider using identification services which are FIDO Alliance certified for best practices:
    - Get Certified for Face Verification | FIDO Alliance - FIDO Alliance
    - Identity Verification Certification Programs | FIDO - FIDO Alliance
    - Battling Deepfakes with Certified Identity Verification | FIDO Alliance - FIDO Alliance
- Include language in job postings stating that reasonable interview accommodations will be provided upon request but the expectation is no audio or video manipulation methods will be allowed during the interview process.
- Educate candidates that are invited to interview that you have processes for identifying candidate impersonators. Also, let them know that you will prosecute all discovered employment fraud. (Sullivan, 2020)
- Implement a series of interviews with different team members and where possible vary the format (video, phone, in-person) and timing of interviews.
- When a candidate is selected for an interview, ensure the process for scheduling that interview includes a disclosure that the interview must be conducted with the camera on, no background blurring or backdrops, no audio or video manipulation or filtering, no headphones, with their screen shared. State again that requests for reasonable accommodations for assistive technology may be made at this point.
- Audit all of your hiring practices to ensure your hiring team is consistently following best practices on background checks, references, resume review, interviews, and more.

Sensitive Data Disclosure

- Review current policy and procedures for sensitive data disclosure which may include mergers and acquisitions, legal, financial transactions, and employee information (HR) disclosures.
- Interview employees in those departments to determine the current workflow (which may be different than documented policy or process).

- Identify gaps in policy, procedure and actual practice.
- Identify processes that permit authorization and authentication through means that are not protected by MFA.
- Inventory the method of human-based authentication, and review for best practices.
- Schedule a review of these procedures on at least an annual basis, as best practices often change according to current threat landscape.

Brand Monitoring

- Complete an inventory of all the departments, tools and services leveraged by the organization for brand and reputation monitoring. Often this monitoring is conducted by multiple teams (CTI, legal brand protection, etc).
- Review monitoring services and platforms to determine if deepfake alerting is within scope.
- Ensure those departments are educated and trained on the procedure for reporting deepfakes.

Event Response

- Identify current mechanisms for reporting deepfakes, any current guidance or awareness for deepfakes.
- Review forensic retainers to determine if digital forensics expertise for deepfake analysis is included, and what the service level agreement (SLA) for that analysis is.
- Determine what service or mechanism your organization uses for takedown requests for look-alike domains and other copyright violations and determine if that service or process is also equipped to handle takedown requests for deepfake content (Gesser et al., 2023).
- Review or establish a deepfake incident response plan.

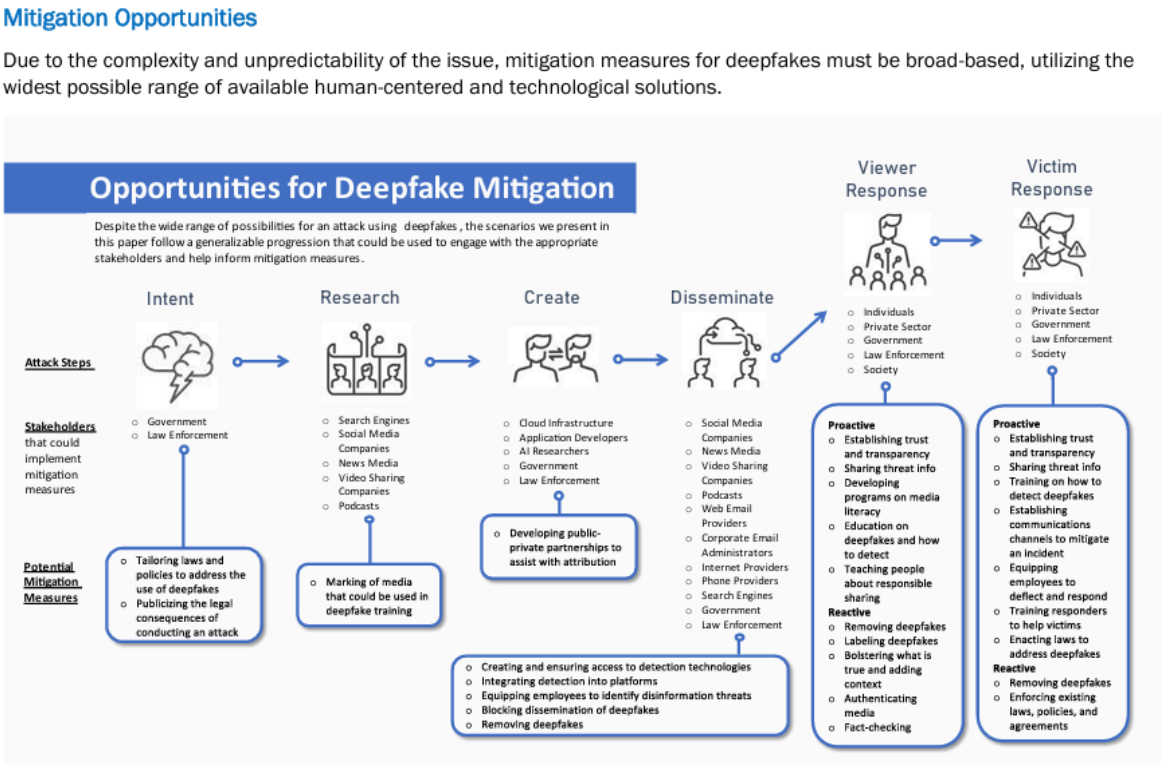
DeepFake Incident Response Plan

Now that you have done the critical homework in your risk assessment phase and understand threats, threat activity, and processes relevant to your specific organization, you can develop a deepfake incident response plan. A deepfake incident response plan is critical as during an event, having clearly outlined roles and responsibilities, templates for communication, and an understanding of how to respond is crucial to a timely response. A quick response helps an organization:

- Mitigate or reduce reputational damage
- Protect sensitive information
- Preserve trust and credibility
- Ensure financial impacts are limited
- Adhere to legal and regulatory compliance requirements
- Ensure operational continuity
- Identify opportunities for mitigation and ensure strategy and processes are in place

A helpful diagram for starting this planning is provided in a document from the Department of Homeland Security called “Increasing Threat of DeepFake Identities.” the opportunities for

mitigation identified:



The deepfake incident response planning process should, at a minimum (Gesser et al., 2022):

- Establish governance structures to oversee security measures and policies related to deepfake threats.
- Document who owns monitoring for deepfakes, what is the alerting process, channels, and stakeholders.
- Document who owns the takedown process for deepfakes, and how escalation is conducted, such as legal action if a takedown request is denied.
- Create a crisis communication plan for each type of the deepfake scenarios described below. In all scenarios, quick and effective communication is key to containment in response.
    - Ensure that templates are developed and approved by all parties with well-defined approval processes for when to implement them.
    - Ensure distribution plans and templates are updated regularly.
- Organizations should consider whether the deepfakes are part of a larger campaign intended to harass, exact revenge, or extort a company or individuals. Incident response plans should account for the following implications (Gesser et al., 2023):
    - Reputational damage.
    - Extortion pressure following a ransomware or data exfiltration event.
    - Hacktivism / corporate activism.
    - Financial fraud.

    - Sensitive information disclosure.
    - Industrial espionage.
    - Computer or network breaches.
    - Misleading stakeholders.
    - Stock prices manipulation.
- Determine if deepfake identification technology needs to be acquired or if existing incident response retainers include this type of forensic analysis. Review the SLA of those retainers and determine if that timeframe is an acceptable delay before making a public statement declaring the content to be verified as fake.
- Define the process and governance for law enforcement involvement.
- Conduct a tabletop exercise to test the deepfake incident response plan. Some specific scenario examples maybe selected from:
    - Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios - Carnegie Endowment for International Peace (https://carnegieendowment.org/research/2020/07/deepfakes-and-synthetic-media-in-the-financial-system-assessing-threat-scenarios)
    - Increasing Threat of Deepfake Identities - DHS (https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf)

Once you have established clear roles and responsibilities, the final stage of preparation is establishing a mechanism for reporting deepfakes and then conducting an education and awareness campaign for all your employees.

Awareness Training

Awareness training for your employees must, at minimum, cover:

- What deepfakes are
- What to do if you think you are being targeted by a deepfake
- What to do if you are a subject of a deepfake
- Where to report deepfakes

There is a significant amount of training material available for deepfake awareness training, which focuses on educating employees to spot indications that audio or video might be fake. However, the authors of this guidance urge you to consider how much of the existing guidance is already outdated and to carefully consider the content of that training before adopting any specific training program.

The majority of current awareness training on deepfakes focus on trying to educate employees how to detect a deepfake, such as providing guidance on detecting imperfections in video such as lip movements or hands, or artifacts in the audio such as unusual pauses. However, this is training individuals to believe it is possible for humans to tell the difference between real and fake in all cases, and as this technology increases in sophistication that guidance will only have served to provide a false sense of expertise and assurance. Researchers have discovered that even with training people both cannot reliably detect deepfakes and tend to overestimate their own ability to identify deepfakes (Köbis et al., 2021). Would it not be preferable, regardless of how perfect the video or audio is, to train employees to follow financial controls and procedures even when pressured to go around them?

Further, deepfakes do not have to be perfect to be effective since scammers typically exploit human psychology by employing urgent and high pressure scenarios which intend to create fear and panic to make targets act rashly. Asking employees who are not experts in audio or video generation to be hyper-aware of potential telltale artifacts and inconsistencies during an intense phone call or high-pressure request seemingly coming from someone in a position of authority seems to be an excessive expectation. When deepfake attacks are successful at conducting fraud or social engineering a target for the purposes of gaining access, these are almost universal examples of when a target was manipulated into bypassing established procedures and not

correlated with the sophistication of the fake. Therefore, defense-in-depth and layered controls is a key strategy to mitigate and prevent the worst impacts of deepfake social engineering.

Additionally, there are a number of benign video call tools which are able to cause the type of artifacts that people typically associate with deepfakes. NVIDIA has several such tools, for example, one feature in the Broadcast tool makes it appear as if the person is always making eye contact with the camera even when they are not. Many of these tools could be disability accommodations or simply quality-of-life improvements, which makes banning them outright unlikely.

Oli Buckley, a professor of cyber security at the University of East Anglia and other experts in the field also recommend that organizations should opt for a change in mindset instead of that approach. "You can't just believe your eyes these days, think a bit more widely about the videos you see, or the calls you get. Critical thinking is the most important factor when dealing with deepfakes, or any scam like this" (Hughes, 2023).

Therefore we recommend your training should:

- Future-proof educational guidance and prevent accidentally conditioning an employee to believe they have the expertise to detect a "real" audio or video. This involves emphasizing that their eyes and ears cannot be trusted. Therefore, the proper process must be followed in all cases without exception.
- Reinforce how deepfakes are designed to influence someone to take an action by triggering strong emotions such as fear and are often pitched with a sense of pressure or urgency. This is referred to as an amygdala hijack, overriding normal logic-based thinking leading a person to take an action before they have time to reflect on the unusual nature of the request (Rowles, 2023).
- Employees should also hear repeated, reinforced guidance on challenging requests from senior leadership to move from authenticated meeting platforms to other conferencing technologies, even when "explained" by "connection issues", or requests which originate from unusual sources such as WhatsApp messages.
- Reinforce with all employees that they are empowered and encouraged to verify all unusual requests by asking for communication through a different channel. Requests via email should be verified with calling a known, good phone number. Requests by video call could be confirmed through an email and so forth.
- Provide employees with guidance on what to do if they suspect a deepfake and where to report it. This could include guidance such as hitting "record" on a conference call, noting any contact details involved (emails, numbers, apps that were used) and making notes about the request (company names used, dollar amounts, bank accounts).
- Standardize and socialize the practice of requiring verification on any meetings before users may join.
- We recommend that your deepfake education approach include similar tactics that are commonly employed for hardening employees against phishing which should include - although often controversial - conducting deepfake simulations to test employee awareness and effectiveness of procedures (Francey, 2024). This should include:
    - Deepfake audio or video impersonation of executives on phone calls or video conferences
    - Deepfake social media profiles and attempt to connect with employees
    - Sending phishing emails or messages with deepfake content to employees

Event Specific Guidance

This section provides specific guidance for Detection and Analysis; Containment, Eradication and Recover and Post-Incident Activity for three most common types of deepfake events.

Financial gain through fraud by impersonation

These events typically involve the impersonation of high-level employees or employees involved in Merger and Acquisition activities (M&A), legal settlement payments, or financial transactions. Threat actors use these deepfakes to trick employees into transfering funds or sharing sensitive information under the guise of highly urgent, extremely confidential requests from company leadership.

Detection and Analysis:

- Review compliance procedures for financial transactions, providing greater latitude to challenge senior leadership requests.
- Review deepfake reporting procedures and educational materials to inform employees involved in financial transactions to make them aware of the reporting mechanism,
- Review the processes for ensuring Separation of Duties and Dual Authorization and determine how automatic identification of any exception to follow this process could be detected.
- Determine the typical rate at which third-parties update their banking or payment information on average in your organization, enabling alerting to detect an unusual number of changes to payment accounts or methods within a limited time frame. For example, if an insurance company expects providers to update their payment account on average 3 times a year, then a change of account three times in a single month should result in an alert and review of the requests.
- Interview the individual who received the deepfake as soon as possible to ask for any details they remember or might have written down about the event and the content of the request. Ask them if they've had any unusual communications leading up to the event, including through social media, personal phone calls or other non-corporate activity.
- Document the incident thoroughly, as this documentation may be critical for future reference including internal reviews, legal proceedings or insurance claims. Ensure the detailed records include names, dates and times that cover (Francey, 2024):
    - Initial discovery of the fraud or attempt
    - All communications with the threat actor
    - Steps taken to notify financial institutions and/or authorities
    - Actions taken to isolate affected systems and accounts

Common TTPs:

- Technique: Gather Victim Information
    - Tactic: Reconnaissance
    - Related ATT&CK TTPs:
        - T1589.003 - Gather Victim Identity Information: Employee Names
        - T1591.004 - Gather Victim Org Information: Identify Roles
        - T1593.001 - Search Open Websites/Domains: Social MediaSearch
        - T1593.002 - Open Websites/Domains: Search Engines
        - T1594 - Search Victim-Owned Websites
    - Description:
        - Adversaries may gather information about the victim's identity that can be used during targeting. Information about identities may include a variety of details like employee names, contact info,

names of divisions/departments, specifics of business operations, relationships, announcements as well as the roles and responsibilities.

In deepfake incidents causing financial fraud, Adversaries typically derive two personas and gather information for both - namely one who has absolute authority to command a transaction and the other who executes the same, the latter typically falls a victim. Victim is often targeted mid-level management or individuals with access to conduct or authorize financial transactions

Adversaries may gather this information in various ways like the ones exposed via online or other accessible data sets(ex: Social Media or Search Victim-Owned Websites).

- Technique: Gather Victim Artifacts
  - Tactic: Reconnaissance
  - Related ATT&CK TTPs:
    - T1593.001 - Search Open Websites/Domains: Social MediaSearch
    - T1593.002 - Open Websites/Domains: Search Engines
    - T1594 - Search Victim-Owned Websites
  - Description:
    Adversaries may gather artifacts of an employee like his genuine images, audio or video clippings that are publicly available which can later be used to create deepfake materials. This will typically be gathered for the persona who is the commanding authority for initiating a financial transaction.

    Adversaries may gather this information from social media, search engines or victim-owned websites.
- Technique: Acquire Non-traceable financial account
  - Tactic: Resource Development
  - Related ATT&CK TTPs:
    - T1583: Acquire Infrastructure
  - Description:
    - Adversaries may create a non-traceable financial account or may use an existing one to collect the funds triggered from the fraudulent financial transaction.
- Technique: Develop Deepfake Models
  - Tactic: Resource Development
  - Related ATT&CK TTPs:
    - T1587.004: Develop Capabilities: Exploits
  - Description:
    Adversaries may create machine learning models to mimic the victim's personal characteristics like their voice, facial expressions using the artifacts collected earlier. They may build these models right from scratch or use AI voice cloning tools, voice changer softwares, Deepfake video tools. They can also subscribe to these softwares as a service to quickly develop the model.

    They would test this model using techniques like Text-to-Speech before it is ready for the real time or offline audio or video clone of the victim.
- Technique: Initiate contact with victim
  - Tactic: Initial Access
  - Related MITRE ATT&CK/ATLAS TTPs:

        - AML.T0052 / T1566 Phishing
        - T1585.001: Establish Accounts: Social Media Accounts
    - Description:
      Contact is initiated through non-standard channels or faked audio or video on an authorized channel but with a request to move the rest of the conversation to a non-standard channel, with a complaint of some technical issue. These include: WhatsApp, LinkedIn instant messaging, text messages, and phone calls (sometimes from spoofed numbers). The voice message is left where the caller appears to have a premade script ready. This often directs the target to then contact or look for follow up on a non-company managed communication channel such as WhatsApp, etc
- Technique: Luring Victim to execute instructions
    - Tactic: Execution
    - Related MITRE ATT&CK TTPs:
        - T1204:User Execution
    - Description:
      Adversaries will attempt to lure the victim to initiate a fraudulent transaction by leveraging urgency on a business requirement. This requirement may be crafted maliciously based on the business announcement & relationships data gathered from the reconnaissance stage. Need for urgency and confidentiality are stressed to the target. This is intended to force the target into bypassing SoD and Dual Authorization processes
- Technique: Evade Existing security controls
    - Tactics: Defense Evasion
    - Related MITRE ATT&CK/ATLAS TTPs
        - T1656: Impersonation
        - T1036: Masquerading
        - T1078: Valid Accounts
        - AML.T0015: Evade ML Model
    - Description:
      Adversaries may create ML models such that the deepfake audio or video produced from it cannot be detected by traditional security solutions. Also, they would use model evasion techniques if an ML-based detection tool is in operation for deepfakes.
- Technique: Deepfake fund diversion & Financial Impact
    - Tactics: Exfiltration & Impact
    - Related MITRE ATT&CK TTPs
        - T1657 - Financial Theft
    - Description:
      Once social engineering is successful, victims can be deceived into sending money to financial accounts controlled by an adversary.Victim would complete the financial transaction diverting the fund requested to an adversary owned non-traceable financial account.

Containment, Eradication and Recovery:

1. If a financial transaction was completed, immediately contact both financial institutions involved and report the fraud. Immediate notification here is key to the potential recovery of funds.
2. If a financial account was provided by the threat actor but a transaction was not completed, notify the financial institution immediately that the account may be being used for fraud.
3. Using information provided by the target of the deepfake request to use keyword searches to look for phishing emails containing similar wording or messages.

4. Do proactive hunting to see if other employees in similar positions of privilege or access are being targeted.
5. Contact the subject of the deepfake, ask them to review their personal accounts and social media for unusual activity, such as other attempts to conduct social engineering, recent fake accounts that may have requested to connect to gain access to media content, evidence of harassment.
6. Investigate the activity of the subject of the deepfake and review any unusual alerts for their account or email.

Post-Incident Activity:

1. Review with the finance department any attempted or fraudulent transactions that were recently requested that involved the targeted deepfake subject or the fraudulent account.
2. Review any statistical aberrations in financial transactions within 90 days of the event including unusually large or small payments for the vendor type, unusually frequent payments, or unusual increase in number of calls or questions on an account.
3. Review any deviations from standard procedures that occurred during this event, determine if there were gaps in alerting or detection mechanisms and improve on them to improve future responses and security measures.
4. Determine if processes for authenticating requests or approval of financial transactions require updating.

Impersonation for cyberattacks

Typically, deepfakes are used to create new accounts and take over existing ones. Threat actors have been reported to use deepfakes to conduct social engineering, bypass authentication or biometric authentication, or further reconnaissance of a target prior to other cyber threat activities.

Detection and Analysis:

1. The detection of these attempts relies almost entirely on the helpdesk being appropriately trained and having a specific reporting mechanism for reporting suspected impersonation events. Review deepfake reporting procedures and educational materials to inform employees involved in financial transactions to make them aware of the reporting mechanism.
2. Create alerts for requests that appear to involve impossible travel by comparing things like geolocation of IP addresses, badge swipes in physical locations, and unusual email routing.
3. Create alerts that compare device fingerprinting for devices commonly seen as used by each user. Logins or requests for new or unrecognized devices should prompt additional verification steps.
4. Consume helpdesk service ticket logs into an analysis platform such as Splunk to create detections for things such as unusual spikes in user requests, large spikes in authentication resets or unusual contact times for known user location.
5. Interview the individual who received the deepfake as soon as possible to ask for any details they remember or might have written down about the event and the content of the request.
6. Document the incident thoroughly, as this documentation may be critical for future reference including internal reviews, legal proceedings or insurance claims. Ensure the detailed records include names, dates and times that cover (Francey, 2024):
    a. Initial discovery of the impersonation or attempt
    b. All communications with the threat actor
    c. Actions taken to isolate affected systems and accounts

Common TTPs:

- Technique: Gather Victim Information
    - Tactic: Reconnaissance
    - Related MITRE ATT&CK/ATLAS TTPs:
        - Refer to the section from "Financial gain through fraud by Impersonation"
        - T1592-Gather Victim Host Information
        - T1590-Gather Victim Network Information
        - T1597-Search Closed Sources
    - Description:
      Refer to the procedure from "Financial gain through fraud by Impersonation" section. In addition to that, adversaries may gather information related to helpdesk and the secret question responses for the persona that they would be using. The information can also be related to the victim environment like the Endpoint Operating systems, Internal network information, Application information so that the adversary is fairly ready to command an instruction for execution.

      Victim can be a help desk falling to someone impersonating a senior executive or it can be an employee falling victim to somebody posing as their superior.

- Technique: Gather Victim Artifacts
    - Refer to "Financial gain through fraud by Impersonation" for details.

- Technique: Develop Deepfake Models
    - Refer to "Financial gain through fraud by Impersonation" for details.

- Technique: Initiate contact with victim
    - Tactic: Initial Access
    - Related MITRE ATT&CK/ATLAS TTPs:
        - AML.T0052 / T1566 Phishing
        - T1585.001: Establish Accounts: Social Media Accounts
    - Description:
      Refer to the procedure from "Financial gain through fraud by Impersonation" section. In addition to that, Adversaries may connect to different service desks of the organization, use the faked audio or video to lure an agent to do something malicious. Ex: Aligning the deepfake content to profile pictures to pass human validation may be a choice here. In other cases, Using non-standard channels, the faked content may be used to mimic a superior trying to get a privileged employee to do something harmful. Technique: Gather Internal technical Information
- Tactic: Discovery
    - Related MITRE ATT&CK/ATLAS TTPs:
        - T1087:Account Discovery
        - T1217:Browser Information Discovery
        - T1652:Device Driver Discovery
        - T1057:Process Discovery
        - T1012:Query Registry
        - T1518:Software Discovery
        - T1082:System Information Discovery
        - T1614:System Location Discovery
        - T1016:System Network Configuration Discovery
        - T1049:System Network Connections Discovery
    - Description:
      Adversaries may use the initial contact to discover more about the system that the victim is using to further craft exploits to get initial access to that system. Mostly, the information is gathered using faked audio or video and by asking questions to the victim or making the victim execute certain commands into the system for discovery. Since the session cannot be kept very long, the adversary may limit doing complicated tasks like the rest of the network discovery for lateral movement at this stage.
- Technique: Initial Access to the victim's systems
    - Tactic: Initial Access
    - Related MITRE ATT&CK/ATLAS TTPs:
        - T1189:Drive-by Compromise
        - T1133:External Remote Services
        - T1200:Hardware Additions
        - T1566:Phishing
        - T1091:Replication Through Removable Media
        - T1078:Valid Accounts
    - Description:
      Adversaries may use an exploit based on the information gathered, use deepfake techniques to make the victim execute the exploit in the system for continuous remote access to the system. The technique could range

from delivering a malicious attachment to an official or personal email address, Tricking the user to click on a malicious link and use drive-by compromise technique or having the user attach a malicious hardware or media device that is shipped to the user earlier.

Note: From here on, adversaries may progress his attack cycle using regular ATT&CK TTPs listed for enterprises.

Containment, Eradication and Recovery:

1. Contact the employee who was impersonated and verify all recent helpdesk requests to identify any invalid requests.
2. Ask the impersonated employee to verify the enrolled devices for Multi-factor Authentication (MFA). If an unidentified device is enrolled, preserve all details from that device for forensic purposes before removing it from Mobile Device Management (MDM) or MFA management systems.
3. Do proactive hunting to see if other employees in similar positions of privilege or access are being targeted. Review helpdesk call logs and tickets within a similar timeframe for unusual trends or clusters based on time, department, region, or type of request.
4. Conduct an OSINT / Social media review of target, this can help identify additional sources for review and put together a timeline of the campaign. When searching, make sure that you have a full list of the details that the impersonator shared to include in your scope. This may include things such as nicknames, preferred names, job titles, department, reporting manager or employee ID which can assist in identifying potential reconnaissance sources for the threat actor. This should include at a minimum OSINT sources for businesses intelligence and sales leads such as.
    - LinkedIn (review the profile, recent posts and comments, who the subject is connected to inside the organization, how their title and department are referenced)
    - Rocketreach.co
    - Lusha.com
    - Uplead.com
    - DNB.com
    - Apollo.io

    As well as other free OSINT tools that can help you discover other social media sites and mentions such as:
    - grep.app
    - OSINT Framework
    - IntelligenceX
    - Social Searcher
    - The Harvester

    The objective of the OSINT exercise is to:
    - Understand where threat actors maybe acquiring social engineering details.
    - Identify other likely targets in your organization for social engineering attempts to conduct further threat hunting.
    - Review of authentication processes in relationship to exposed information about employees to identify weaknesses. For example, if an employee ID is exposed on a github, then it should not be used for authentication.
5. Determine if the target's credentials were a part of any recent data leak or breach (stealer logs, haveibeenpwned, etc). If a stealer log is discovered, review the entire entry for the stealer log, including the autofill information that may provide advanced social engineering opportunities or answers to challenge questions. Notify impersonated employees of any findings so they may secure non-corporate accounts that may use these details to perform password resets.

6. Review any identification provided for possible synthetic identity. Notify the impersonated employee of any findings.
7. Review accounts for attempted extortion, threats, harassment, or disclosure of sensitive details that could lead to impersonators successfully authenticating or further social engineering

Post-Incident Activity:

1. Consider referencing the same sources when producing your next deepfake simulation for employee awareness training.
2. Conduct post-incident reviews to identify areas for improvement in security controls and response procedures.

Job Interview Fraud

Recent campaigns, particularly those attributed to the Democratic People's Republic of Korea (DPRK), have highlighted a growing threat in the recruitment process: the use of deepfake technology in job interviews. Malicious actors are leveraging this technology to secure positions within organizations, potentially gaining insider access for further exploitation. This guide aims to provide strategies for prevention, detection, and investigation of such incidents, with a focus on methods that do not rely solely on detecting fake audio or video.

These efforts appear to serve multiple purposes:

1. Intelligence gathering: Gaining insider access to valuable information and technologies.
2. Financial gain: Manipulating cryptocurrency markets or facilitating cyber heists.
3. Deployment of malware: Gaining access to internal networks and systems as an employee allows the actors to deploy malware to maintain persistence even if they are terminated and perform exfiltration.

These tactics highlight the need for strong identity verification in remote hiring, especially for sensitive positions. However, it's important to balance security with accessibility. Some legitimate candidates may use assistive technologies during interviews as an accommodation for disabilities.

For instance, NVIDIA Broadcast, a free software for users with newer NVIDIA GPUs, offers an "eye contact" feature. This tool is popular in neurodivergent communities as an aid for interviews.

Employers must therefore strike a balance between preventing fraud and ensuring fair access for all candidates, including those with disabilities who may benefit from assistive technologies.

Detection and Analysis:

In 2023 report, NISOS Investigators found the following commonalities in the personas' profiles and resumes:

- Personas claim to have experience developing web and mobile applications, knowledge of multiple programming languages, and an understanding of blockchain technology.
- Personas have accounts on employment and people information websites as well as IT industry-specific freelance contracting platforms, software development tools and platforms, and common messaging applications, but typically lack social media accounts, suggesting that the personas are created solely for the purpose of acquiring employment.
- Photos of the same individual are used to create multiple personas.
- Personas have several accounts with the same name and photo that are sometimes associated with different locations, some of which are abroad.
- Personas' accounts contain only minimal information, and some of the resume content on the accounts is likely copied from real individuals in the IT industry.

Additional measures for detection and analysis:

- Ensure HR and interviewers are educated about the trend of interview fraud, and know how to report their concerns to cybersecurity.
- Consider implementing risk scoring based on multiple factors to flag potential fraud.
- Ensure there is a process for cybersecurity to be able to review the Applicant Tracking Systems for emails or patterns in resumes which match known malicious activity. Many ISACs now are sharing indicators regarding job fraud attempts which will be only available in applicant systems, which are not typically centralized for logging and monitoring.

- Foster a culture of awareness and skepticism about hiring for remote roles. Be sure and educate your hiring team. Make them aware that applicants with major flaws or lower qualifications may be fraudulently impersonating a candidate. If an applicant appears to be "too good to be true" for your job and company, it's okay to assume the worst and heighten your screening effort (Sullivan, 2020).
- Consider requiring at least one in-person interview for remote roles; even the mention of an in-person interview requirement maybe enough to deter fraud.
- When conducting interviews and following the guidance regarding virtual interview must be conducted with the camera on, no background blurring or backdrops, no audio or video manipulation or filtering, no headphones, with their screen shared. This is to detect if there are other people in the room also interviewing, or if there is visible cameras in the room with the candidate, as they are often monitored by the DPRK.
- Conduct thorough background checks, including employment history and educational qualifications.
- Verify references and previous employers personally.
- Ask for additional verification of address. Send equipment only to the address on the ID where possible. If this is not possible, ship the equipment to a facility that will require them to physically who up and present ID to obtain their company laptop. Consider requiring the employee to sign a document which had language warning them that receiving this equipment means they acknowledge that providing access to that equipment to others is s federal offense and the organization will prosecute to the fullest.
- Start new employees in a highly restricted environment where they only have access to the systems necessary to perform their work. Ensure they don't have immediate access to production systems or sensitive data. Gradually increase access based on job requirements and demonstrated trustworthiness (Fisher Philips, 2024) .
- Implement mandatory probationary periods for new hires.
- Consider the use of User and Entity Behavior Analytics (UEBA) tools to detect anomalous behavior.
- Secure all related digital artifacts, including interview recordings, email communications, and application materials.
    - Analyze metadata of submitted documents and interview recordings.
    - Look for signs of manipulation or inconsistencies in file properties.
    - Analyze digital footprints and online presence for consistency, including reverse image searches, doing exact word matches from the resume or application
    - Cross-validate identity information across multiple data sources including professional social media
    - Some defenders have recommended using pimeyes.com during investigation
    - Look for files using the languages that are not listed as the primary or secondary language for the employee, such as guides in Korean that provide guides on how to obtain documentation in other countries, such as how to get a driver's license in another country. Actors are known the exchange such guides.
    - Many DPRK workers also have profiles on Guru.com and will review each other to add legitimacy (NISOS, 2023).
- Review logs of any remote access or VPN connections used during interviews.
- Perform traceroutes to look for two-hop connections.
- Examine phone numbers to determine if they are VoIP.
- Review network traffic for unusual dual VPN connections or KVM or IP activity (e.g. 5900 for VNC). Review system logs for KVM-related events or connections.

Remember that many employees will use KVMs for legitimate reasons, so consider this just one element of the investigation.
- Check for unusual port forwarding rules
- Monitor for an unusually high number of changes to the direct deposit account an employee uses for payroll; many of these workers may rely on money mules or accounts that may get shut down when fraud is suspected.
- Review active network connections.
- Analyze bandwidth usage patterns that may indicate video streaming.
- Analyze IP addresses and geolocation data for inconsistencies. (e.g. using Spur.us)

Common TTPs:

- Technique: Initial Access
  - Related ATT&CK TTPs:
    - T1199: Trusted Relationship
    - T1078: Valid Accounts
    - T1589: Gather Victim Identity Information
    - T1070.001: Indicator Removal on Host, File Deletion.
  - Description:
    Adversaries may use a variety of techniques in order to gain access to an interview process. For example, they could use stolen LinkedIn credentials to apply for jobs. Additionally, they could collect data about individuals or organizations to create believable fake identities or to facilitate targeted interview attempts.
- Technique: Execution
  - Related ATT&CK TTPs:
    - T1059: Command and Scripting Interpreter
  - Description:
    The adversary could abuse command and script interpreters to connect back to adversary controlled networks.
- Technique: Persistence
  - Related ATT&CK TTPs:
    - T1136: Create Account
    - T1543: Create or Modify System Process Create or Modify System Process
  - Description:
    Adversaries could attempt to create additional accounts in order to maintain access to a company issued laptop.
- Technique: Privilege Escalation
  - Related ATT&CK TTPs:
    - T1484: Domain Policy Modification
- Technique: Defense Evasion
  - Related ATT&CK TTPs:
    - T1036: Masquerading
    - T1550: Use Alternate Authentication Material
    - T1565: Data Manipulation
    - T1027: Obfuscated Files or Information - Transferring potentially harmful files using a Raspberry Pi may involve obfuscating the files to avoid detection by security systems.
- Technique: Discovery
  - Related ATT&CK TTPs:
    - T1087: Account Discovery

        - T1082: System Information Discovery
- Technique: Lateral Movement
    - Related ATT&CK TTPs:
        - T1021: Remote Services
- Technique: Collection
    - Related ATT&CK TTPs:
        - T1114: Email Collection
        - T1213: Data from Information Repositories
        - T1005: Data from Local System
- Technique: Exfiltration
    - Related ATT&CK TTPs:
        - T1048: Exfiltration Over Alternative Protocol
        - T1052.003 Exfiltration Over Physical Medium: Removable Media - Transferring files using a Raspberry Pi indicates the use of removable media to exfiltrate data from the organization.
- Impact
    - Related ATT&CK TTPs:
        - T1499: Endpoint Denial of Service
        - T1565: Data Manipulation
        - T1486: Data Encrypted for Impact
        - T1490: Inhibit System Recovery
        - T1005: Data from Local System
        - T1056: Input Capture

Containment, Eradication and Recovery:

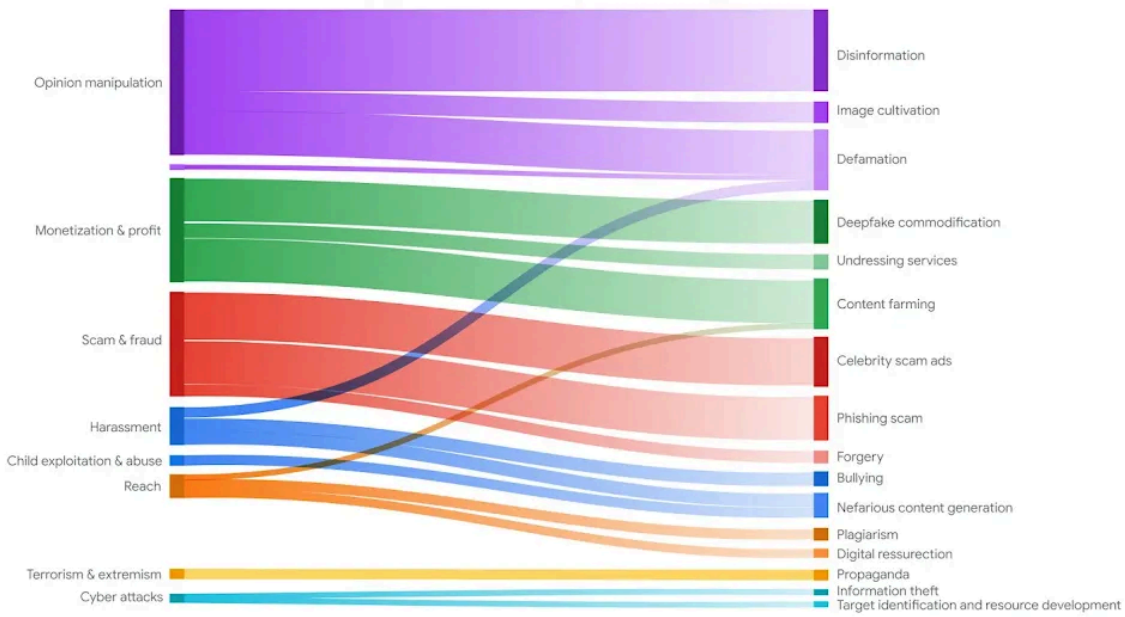
- Consult with legal counsel to ensure compliance with privacy laws and employment regulations.
- If a breach is confirmed, activate your incident response plan.
- Isolate affected systems and revoke any access granted to the malicious actor.

Post-Incident Activity:

- Create a detailed timeline of the hiring process, from initial application to the report of suspicion.
- Collect any bank accounts or other financial accounts used by the suspect
- Identify any anomalies or deviations from standard procedures.
- Report the incident to relevant authorities and information sharing organizations.
- Share sanitized details with industry peers to raise awareness and improve collective defense.

Mis/Dis/Mal Information

According to a comprehensive study conducted by Google DeepMind, mis/dis/mal information is the leading way in which malicious actors abuse Generative AI. Threat actors motivated by political ideology, such as hacktivists, could leverage generative AI to defame a company or government by creating fake representations of organizational leaders saying offensive things. Similarly, criminals could seek to manipulate the stock price of a company by creating a deepfake of a CEO making a major business announcement (Reuters, 2024).



(Graphic via Google DeepMind)

Detection and Analysis

Detection efforts typically focus on developing methods that seek evidence of manipulation and present that evidence as a numerical output or a visualization to alert an analyst that the media needs further analysis. These methods are developed assuming that modifications to the original media or completely synthetic media contain statistically significant traces that can be found. This form of detection is a cat-and-mouse game; as detection methods are developed and made public, there is often a quick response from the generation community to counter them. (media.defense.gov)

For several years, public and private organizations have expressed concern over manipulated multimedia and developed means to detect and identify countermeasures. Many public and private partnerships have since emerged, focusing on cooperative efforts to detect these manipulations and verify/authenticate multimedia (DOD, 2023).

A 2024 assessment of deepfake detection challenges by GAO point out several limitations to the current maturity of detection technology including:

- Data needs for detection. Deepfake detection tools must generally be trained with large and diverse data sets to reliably detect deepfakes. Technology companies and researchers have released data sets to help train detection tools, but the current data sets are not sufficient by themselves. Detection tools must be constantly updated with data of increasing sophistication to ensure that they continue to be effective at detecting manipulated media.

- Cat and Mouse - techniques used to identify deepfakes tend to lead to the development of more sophisticated deepfake techniques. This "cat and mouse" situation means detection tools must be regularly updated to keep pace
- According to recent studies, existing detection methods and models may not accurately identify deepfakes in real-world scenarios. For example, accuracy may be reduced if lighting conditions, facial expressions, or video or audio quality are different from the data used to train the detection model, or if the deepfake was created using a different method than that used in the training data. Further, future advances in deepfake generation are expected to eliminate hallmarks of current deepfakes, such as abnormal eye blinking (GAO, 2024).

While there are some recent advancements in the detection technology space, such as Intel's FakeCatcher (McFarland, 2024) that show promising initial results, our belief is that the maturity of the technology offerings in the space is limited and due to the challenges above, will require substantial investment and ongoing research into accuracy over time before most organizations should consider adoption. For now, the audience for the acquisition of this type of bleeding edge deepfake detection solutions remains the government, defensive agencies and media organizations.

Our recommendations:

- Encourage employees to report video or audio of organizational leaders being released through non-standard channels or making unusual claims.
- Leverage reputation and brand monitoring services to detect when there are spikes in negative sentiment against your organization. Work with your vendor or teams responsible for monitoring sentiment around important times such as earnings announcements.
- If possible, secure the original or a copy of the deepfake content.
- If you have a forensic or incident response retainer that included deepfake analysis, provide the original file and its hash value to forensic examiners.
- If you do not have a retainer for analysis, examine metadata looking for inconsistencies such as timestamps, location or tools used to generate the media. Various guidance on how to do this type of analysis can be found at WITNESS Media Lab | WITNESS Media Lab Verification Resources. Variations in recording parameters, mismatched metadata or jumps in timestamps can indicate the media has been manipulated. Some suggestions for this type of analysis include:
  - InVID Verification Plugin - InVID project (invid-project.eu)
  - Video / Photo / Audio tools Digital Journalism | OSINT Essentials
  - Collection of tools for Video Editing and Analyzing can be found at cipher387/osint_stuff_tool_collection: A collection of several hundred online tools for OSINT (github.com)
- Through OSINT, look for additional sources of the media. This can help determine the provenance of the data and additional metadata may be found in other sources. It can also be used to identify if any other internal resources are a target of a campaign. Some suggestions for this include:
  - TinEye and Google Reverse Image Search
  - Collection of tools for Image Search and Identification can be found at cipher387/osint_stuff_tool_collection: A collection of several hundred online tools for OSINT (github.com)

Common TTPs:

- Technique: Gather Victim Information
  - Tactic: Reconnaissance
  - Related MITRE ATT&CK/ATLAS TTPs:
    - Refer to the section from "Financial gain through fraud by Impersonation"

    - Description:
      Refer to the procedure from "Financial gain through fraud by Impersonation" section. In this case the victim could just be one persona as opposed to multiple like what is in previous events.
- Technique: Gather Victim Artifacts
    - Refer to "Financial gain through fraud by Impersonation" for details.
- Technique: Finalize hosting options
    - Tactic: Resource Development
    - Related MITRE ATT&CK/ATLAS TTPs:
        - T1583.001:Acquire Infrastructure: Domains
        - T1583.006: Acquire Infrastructure: Web Services
        - T1583.008:Acquire Infrastructure: Malvertising
        - T1585.001: Establish Accounts: Social Media Accounts
    - Description:
      Adversaries would look to acquire infrastructure or create fake profiles to upload & publish faked audios or videos. The infrastructure could be a dedicated website, as an advertisement or video content in existing websites or via social media like faked linkedIn profiles, youtube accounts, whatsapp accounts etc.
- Technique: Develop Deepfake Models
    - Refer to "Financial gain through fraud by Impersonation" for details.

- Technique: Upload Deepfake Materials
    - Tactic: Initial Access
    - Related MITRE ATT&CK/ATLAS TTPs:
        - No current TTP mapping
    - Description:
      Adversaries upload the faked materials into the hosted environment finalized in the previous TTP.
- Technique: Attempt Mass Circulation
    - Tactic: Lateral Movement
    - Related MITRE ATT&CK/ATLAS TTPs:
        - No current TTP mapping
    - Description:
      Adversaries may attempt to initiate mass circulation of the content. This could be achieved via posting the faked audio or video directly from a fake social media profile (Like LinkedIn, Whatsapp, Youtube etc) or as a link to the content which is hosted in an adversary controlled infrastructure.
- Technique: Reputational Harm
    - Tactic: Impact
    - Related MITRE ATT&CK/ATLAS TTPs:
        - AML.T0048.001: External Harms: Reputational Harm
    - Description:
      Reputational harm involves a degradation of public perception and trust in organizations. Examples of reputation-harming incidents include scandals or false impersonations.

Containment, Eradication and Recovery:

- Takedown Requests: if the deepfake contains any copyrighted material it maybe possible to submit a "takedown notice" under the Digital Millennium Copyright Act (DMCA) to the website on which the infringing deepfake is hosted. However, it should be noted that this could be fraught with issues if it is not clear there was copyright

infringement or if there is an argument for "fair use" considerations (Gesser et al., 2023).
- Terms of Service Violation Reporting: another possible method for removing deepfakes involves reviewing the hosting website's terms and conditions to determine if manipulated or synthetic media clauses may allow you to submit a violation report (Gesser et al., 2023).
- Establish contacts with members of your organization's public relations/communications team.

Post-Incident Activity:

- Conduct post-incident reviews to identify areas for improvement in security controls and response procedures.
- Determine if additional training is needed for your staff. Your people are your best resource and an educated workforce will reduce your risk. Everybody, no matter their role in your company, has a responsibility for the security and privacy of your data as well as that of your customers.
- Evaluate if deepfake detection technologies should be acquired, know that these are still immature (GAO, 2024) as discussed above.

Conclusion

As we've explored throughout this guide, the landscape of deepfake technology is rapidly evolving. Both the generation and detection of deepfakes are advancing at a remarkable pace, presenting a dynamic challenge for cybersecurity professionals and organizations alike. While it's impossible to predict with certainty how these technologies will develop, the guidance provided in this document has been crafted with longevity in mind.

Our approach focuses on fundamental principles and best practices that are likely to remain relevant regardless of technological advancements. By emphasizing critical thinking, robust authentication processes, and layered security controls, we've aimed to create a framework that can adapt to future developments in deepfake technology.

Key aspects of our guidance, such as:

- Focusing on process adherence rather than visual or auditory detection of fakes
- Implementing and maintaining strong financial controls and verification procedures
- Cultivating a culture of awareness and skepticism towards unusual requests
- Developing and regularly updating incident response plans

These elements are designed to be resilient against evolving threats. They don't rely on specific technological solutions that may become obsolete but instead build on fundamental security principles and human vigilance.

As deepfake technologies continue to advance, it will be crucial for organizations to stay informed about new developments. However, by implementing the strategies outlined in this guide, organizations can build a strong foundation for defending against deepfake-related threats, both now and in the future.

Remember, the most effective defense against deepfakes isn't just about having the latest detection technology—it's about creating an environment where deception is difficult to achieve, regardless of how convincing the fake may appear. By focusing on these enduring principles, organizations can maintain resilience in the face of evolving deepfake threats.

References

If a cybersecurity firm can fall for the latest AI workplace scam, so can you: 10 steps to protect your business. Fisher Phillips. (2024, August 1). https://www.fisherphillips.com/en/news-insights/latest-ai-workplace-scam-10-steps-to-protect-your-business.html

Bateman, J. (2020, June 8). Deepfakes and Synthetic Media in the financial system: Assessing threat scenarios. Carnegie Endowment for International Peace. https://carnegieendowment.org/2020/07/08/deepfakes-and-synthetic-media-in-financial-system-assessing-threat-scenarios-pub-82237

Brooks, T., G., P., Heatley, J., J, J., Kim, S., M, S., Parks, S., Reardon, M., Rohrbacher, H., Sahin, B., S, S., S, J., T, O., & V, R. (2022). Increasing Threat of Deepfake Identities. Department of Homeland Security.

Chen, H., & Magramo, K. (2024, February 4). Finance worker pays out $25 million after video call with Deepfake "chief financial officer." CNN. https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html

Chickowski, E. (2023, December 8). Criminals use deepfake videos to interview for remote work. Dark Reading. https://www.darkreading.com/cyberattacks-data-breaches/criminals-deepfake-video-interview-remote-work

Ciancaglini , V., & Sancho, D. (2024, May 8). Back to the hype: An update on how cybercriminals are using genai. Trend Micro. https://www.trendmicro.com/vinfo/gb/security/news/cybercrime-and-digital-threats/back-to-the-hype-an-update-on-how-cybercriminals-are-using-genai

Cox, J. (2023, February 23). How I broke into a bank account with an AI-generated voice. VICE. https://www.vice.com/en/article/dy7axa/how-i-broke-into-a-bank-account-with-an-ai-generated-voice

Deloitte. (2024, March). How to safeguard against the menace of deepfake ... https://www2.deloitte.com/content/dam/Deloitte/in/Documents/risk/in-ra-safeguarding-against-deepfake-technology-noexp.pdf

DOD. (2023, September 12). Contextualizing Deepfake Threats to Organizations . https://media.defense.gov/2023/Sep/12/2003298925/-1/-1/0/CSI-DEEPFAKE-THREATS.PDF

EIN Presswire (2024, April 2). It only takes 35 seconds to create a deepfake video/photo-deepfake tools study by humanorai.io. KGET 17. https://www.kget.com/business/press-releases/ein-presswire/700218667/it-only-takes-35-seconds-to-create-a-deepfake-video-photo-deepfake-tools-study-by-humanorai-io/

Francey, E. (2024, July 8). CEO fraud turbocharged by deepfake. Breacher.ai. https://breacher.ai/uncategorized/ceo-fraud/

GAO. (2024, March 11). Science & Tech spotlight: Combating deepfakes. Science & Tech Spotlight: Combating Deepfakes. https://www.gao.gov/products/gao-24-107292

Gesser, A., Bannigan, M., Ford, C. S., Gressel, A., & Caravello, S. M. (2023, January 24). Responding to malicious corporate deepfakes. Debevoise Data Blog. https://www.debevoisedatablog.com/2023/01/24/responding-to-malicious-corporate-deepfakes/

Gesser, A., Gressel, A., Roberts, M. R., Goldstein, C., & Rubinstein, E. (2022, April 27). The value of Ai incident response plans and tabletop exercises. Debevoise Data Blog. https://www.debevoisedatablog.com/2022/04/27/the-value-of-airps-and-ai-tabletops/

How to defend your company against Deepfake Scams. Coro Cybersecurity. (2024, February 20). https://www.coro.net/blog/how-to-defend-your-company-against-deepfake-scams

Hughes, A. (2023, August 26). Ai: Why the next call from your family could be a deepfake scammer. BBC Science Focus Magazine. https://www.sciencefocus.com/future-technology/ai-deepfake-scam-calls

Krietzberg, I. (2023, May 22). S&P sheds $500 billion from fake Pentagon Explosion. The Street. https://www.thestreet.com/technology/s-p-sheds-500-billion-from-fake-pentagon-explosion

Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. iScience, 24(11), 103364. https://doi.org/10.1016/j.isci.2021.103364

Marchal, N., & Xu, R. (2024, August 2). Mapping the misuse of Generative AI. Google DeepMind. https://deepmind.google/discover/blog/mapping-the-misuse-of-generative-ai/

McFarland, A. (2024, August 1). 5 best deepfake detector tools & techniques (August 2024). Unite.AI. https://www.unite.ai/best-deepfake-detector-tools-and-techniques/

Nelson, A., Rekhi, S., Souppaya, M., & Scarfone, K. (2024). Incident response recommendations and considerations for Cybersecurity Risk Management: National Institute of Standards. https://doi.org/10.6028/nist.sp.800-61r3.ipd

NISOS. (2023). (rep.). Investigation: Probable DPRK Online Personas Used to Fraudulently Obtain Remote Employment at U.S. Companies . NISOS. https://6068438.fs1.hubspotusercontent-na1.net/hubfs/6068438/dprk-it-worker-scam.pdf

Onfido. (2024, April 22). Identity fraud insights report 2024. https://onfido.com/landing/identity-fraud-report/

Reuters. (2024, April 10). Beware of deepfake of CEO recommending stocks, says India's National Stock Exchange. https://www.reuters.com/technology/cybersecurity/beware-deepfake-ceo-recommending-stocks-says-indias-national-stock-exchange-2024-04-10/

Rowles, R. (2023, September 12). Amygdala hijacking and social engineering. Security Through Education. https://www.social-engineer.org/social-engineering/amygdala-hijacking-and-social-engineering/

Sullivan, J. (2020, July 20). Identify fraud with remote hiring – could your new-hire be an impersonator?. Dr John Sullivan. https://drjohnsullivan.com/articles/identify-fraud-with-remote-hiring-could-your-new-hire-be-an-impersonator/

Tummalapenta, S. (2024, July 29). How a new wave of deepfake-driven cyber crime targets businesses. Security Intelligence. https://securityintelligence.com/posts/new-wave-deepfake-cybercrime/

VandeHei, J., & Allen, M. (2023, November 8). Behind the curtain: What ai architects fear most (in 2024). Axios. https://www.axios.com/2023/11/08/ai-fears-deepfake-misinformation

NSA, FBI & CISA (2023, September 12). Cybersecurity Information Sheet on Deepfake Threats. CISA. https://www.cisa.gov/news-events/alerts/2023/09/12/nsa-fbi-and-cisa-release-cybersecurity-information-sheet-deepfake-threats