FS-ISAC

Deepfakes in the Financial Sector: Understanding the Threats, Managing the Risks

A Report by the FS-ISAC
Artificial Intelligence Risk
Working Group

October 2024

Contents

Executive Summary 3

Introduction 3

Understanding Deepfakes in the Financial Industry Context 5

Assets Under the Threat of Deepfakes 6

Deepfake Impersonation Victims 7

Deepfake Recipients 8

Deepfake Taxonomy: Threats to Institutions and Technologies 8

Summary of Controls to Protect Against Deepfake Attacks 13

Conclusion 15

Executive Summary

Deepfake video, audio, or images are a potent emerging threat to financial services firms. These falsified media amplify the impact of disinformation campaigns, enable sophisticated social engineering crimes, and undermine society's trust in the financial system.

What's a Deepfake?

A portmanteau of "deep learning" and "fake," **deepfakes** are synthetic media produced by GenAI. Deepfake videos replace a person's likeness with someone else's in existing images or videos, for example, or mimic a person's voice in deepfaked audio with striking accuracy.

To help financial institutions defend themselves from this new attack vector, the members of FS-ISAC's AI Risk Working Group developed a Deepfake Taxonomy for the financial sector, the first of its kind to define and detail the threat that deepfakes pose specifically to financial institutions.

This document is engineered to help business leaders identify their risks. It covers:

- How financial services institutions are likely to be attacked by deepfakes
- The assets threatened by deepfakes
- The primary recipients of deepfakes
- A summary of controls available to financial firms

The work is timely. Powerful generative artificial intelligence (GenAI) tools make it easy for even low-skill threat actors to create deepfakes aimed at consumers, finance leaders, and the public. And today's fractured media environment, political polarization, and geopolitical tensions facilitate deepfake crimes.

We hope that by sharing this information, cybersecurity teams can enact preventative measures and control mitigations to protect their firms, customers, and reputations, and the public's trust in the financial system.

Introduction

The financial sector stands at the forefront of technological innovation, constantly adapting to new digital landscapes by enhancing the customer experience, streamlining operations, and maintaining security. However, with each technological advancement comes new challenges, and perhaps none is more pressing or potentially disruptive than the rise of deepfake technology.

1 in 10

Companies have encountered deepfake fraud

6 in 10

Executives say their firms have no protocols regarding deepfake risks[1]

Though benign and helpful applications exist, threat actors use deepfakes to bypass traditional security measures, exploiting the human element of trust that often underpins financial transactions and decision-making processes. The function of adversarial deepfakes includes impersonating customers, employees, public officials, and institutional leaders with purposes such as committing fraud and manipulating markets through stakeholder and public deception.

To help counter the threat deepfakes pose, the FS-ISAC Artificial Intelligence (AI) Risk Working Group researched, analyzed, and taxonomized the threats posed by deepfakes, presented here for senior executives and board members who are not necessarily security experts. Upcoming white papers for technologists will discuss controls and mitigations.

This white paper provides broad categories and a common language of deepfake threats and controls to counter them. We distinguish between the different types of threats facing financial services institutions — rather than lumping all deepfakes into a single category of threat – to help firms identify which threat categories pose the greatest risks to their institution so they can focus on the controls specific to those risks.

> Losses from deepfake and other AI-generated frauds are expected to reach tens of billions of dollars in the next few years.[2]

We organize the white paper as follows:

- Deepfake scenarios that level-set the threat financial services firms face.
- A description of the types of deepfake risks according to those scenarios.
- Assets, victims, and recipients threatened by deepfake technology.
- The Taxonomy, which consists of nine classes across two domains.
- A summary of controls that can help address the threats identified in the Taxonomy.

Types of Deepfake Risks to Financial Services Institutions

Financial institutions face significant risks from deepfakes, including:

- **Market risk:** False information can be used to influence and manipulate financial markets.
- **Information security risk:** A deepfake attack can be the first step in enabling malicious actors to infiltrate the institution's systems and exfiltrate sensitive data.
- **Fraud risk:** Deepfakes as a component of social engineering can be used to defraud the institution's customers and clients, or to target the financial institution itself in a fraud scheme.
- **Regulatory risk:** Hiring or supplying sanctioned individuals may be illegal, even if the candidate's identity was deepfaked, and could present legal and regulatory risks to the financial services institution.
- **Reputational risk:** Disinformation campaigns leveraging deepfakes can damage consumer trust and the institution's credibility, causing long-term damage to its brand and reputation.

Understanding Deepfakes in the Financial Industry Context

Figure 1, below, provides a few scenarios demonstrating the risks from deepfakes to financial institutions. Many of these attacks have already been observed within the sector and in non-financial contexts, as the examples show.

Imposter Target	Victim Set	Scenario	Purpose	Example
C-suite impersonation	Banks, investors, the public	A deepfake video of a prominent bank created as a part of an information operations campaign and posted to social media.	Sow confusion, move markets, conduct pump and dump schemes.	A Pentagon explosion deepfake sent the Dow Jones down 85 points in four minutes. (May 2023)
C-suite impersonation	Employees in finance, asset/wealth management, IT/Helpdesk operations, etc.	An audio deepfake enables the impersonation of a C-suite executive to initiate a transaction, transfer funds, or request access to non-public information.	Fraud, access, or to obtain sensitive information, company secrets, PII, and other non-public data.	CFO impersonation in Zoom meeting resulting in transfer of $25 million to fraudster. (February 2024)
Banking consumers	Banking consumers	Banks allowing voice authentication without additional measures enables fraudsters with voice sample sets and stolen PII to initiate transfers, cash outs, and other illicit transactions.	Consumer fraud	WSJ reporter cloned herself with AI and fooled her bank (April 2023)
Financial advisors/ investment bankers	General public, consumers	A deepfake video used to impersonate financial advisors, investment bankers or other trusted advisors to commit fraud.	Fraud	Sydney, Australia financial advisor deep faked himself for client interactions. (July 2023)
Third-party trusted relationships	Financial institution	Threat actors use deepfake impersonations of financial services and external-entity employees to gain access to or exfiltrate money from financial services firms.	Fraud	Deepfake BEC of company director resulted in $35 million transfer by Hong Kong bank manager. Deepfake phone call was followed by emails from bank manager and lawyer. (2020)

Consumers	Financial institution	Fraudster uses GenAI to create fake driver's license and other documents to set up bank accounts.	Fraud, laundering	Underground website that generates $15 fake identification sufficient to bypass verification systems. (February 2024)
Individuals/ general public	Financial services institutions	Threat actors use GenAI to create fake identification documents to bypass HR checks and gain employment for espionage, sanctions avoidance, or malicious activity.	Espionage, revenue generation, sanctions avoidance, initial access	Cybersecurity firm hired a North Korean IT worker using a deepfake identity to obtain employment as a AI software engineer. (July 2024)
Public personas	C-suite	A deepfake of a public persona used to fool a bank's leadership or C-suite to cause embarrassment and discredit their security and vetting procedures.	Embarrassment	ECB President Lagarde and US Fed Chair Powell tricked with deepfakes of Ukrainian President Zelensky in phone calls. The fraudster released clips of their private statements. (Reported April 2023)

Figure 1. Sample deepfake scenarios.

Assets Under the Threat of Deepfakes

Financial institutions often have several classes of assets threatened by deepfake attacks. Some assets can be spoofed by deepfakes, others can be manipulated to make deepfake attacks more successful. The assets below are frequently the subjects of deepfake attacks, and thus require the highest degree of attention.

C-suite media, which encompasses images, audio recordings, and video recordings of high-level executives. These media forms are essential for a variety of purposes, including internal communications and external public relations.

Customer biometrics including data for identity verification, biometric authentication, facial recognition, liveness detection, and fraud indicators. Customer biometrics are crucial for ensuring the security and authenticity of customer interactions and transactions.

Deepfake detection models can be manipulated to remove obstacles to systems and information. Infrastructure hardened to address the threat of deepfakes should contain an array of tools designed to detect fraudulent media.

Infrastructure components are a target because attackers must compromise infrastructure to mount certain attacks. Such components support the entire system with model inference APIs, API credentials, and data stores. These data stores are designated for C-suite media, customer biometric data, and various model training datasets.

Targets of Deepfake Impersonation: Financial Institution Employees and External Entities

Threat actors use deepfake impersonations of financial services and external entity employees to gain access to or exfiltrate money from financial services firms.

- Privileged employees: those with access to systems and responsibility within the organization.
  - > Administrative accounts
  - > Business privileged users
  - > Executives who have access to restricted information and are entitled to perform actions useful to cybercriminals
- Non-privileged employees: those who could enable adversarial exploits through their actions or divulged information.
- External entities: those whose perceived behavior influences others.
  - > Financial firm consumers and clients
  - > Third-party employees
  - > Public figures

Deepfake Impersonation Victims

The Deepfake Taxonomy provides context to describe how threats may impact an organization. We begin with a discussion on impersonation targets – individuals likely to be the subject of deepfake imitation – categorized as financial institution employees and external entities.

Internally, both privileged employees and non-privileged employees can be at risk of impersonation.

Privileged employees have access to systems and information useful to attackers. Executives in particular are vulnerable to impersonation, as they have broad authority and feature in a significant amount of publicly available media that adversaries use to create deepfakes.

Non-privileged employees have less authority but impersonating them allows adversaries to manipulate other staff members (often by creating a sense of urgency) into performing actions that enable the cybercrime. Non-privileged targets have less media presence than senior executives, so their images and voices are typically captured via phone calls or online meetings.

External entities may be impersonated to permit the threat actor access to a financial firm's information or systems, to provoke a response that manipulates markets, or to gain access to individual accounts.

Recognizing attackers' patterns allows institutions to better prepare and defend against the misuse of deepfake technology. Understanding these targets helps firms recognize victims and indicates likely sources of data collection for deepfake creation.

Deepfake Recipients

Deepfake social engineering attacks have two victims: the individual who is impersonated in a deepfake, and the individual who receives the deepfake.

Employees of financial institutions and their external partners may receive a deepfake of an authority in the financial firm, such as a CFO, with a logical reason to perform a proscribed act, such as transferring funds. HR functions may receive deepfakes of job candidates, notably IT workers, created by nation-state threat actors to disguise the identity of their own operatives intent on stealing information, spying on the institution, and disrupting business.

Internally, both privileged and non-privileged employees are potential deepfake recipients, but their vulnerabilities differ. Deepfakes targeting non-privileged employees, such as those in customer support, rely on the trust employees have in their workplace directives, though privileged employees are a higher value target due to their authority and internal access/permissions.

Individuals Likely to Receive Deepfakes in a Social Engineering Attack

Employees of financial services firms and external vendors can be tricked by impersonations received in social engineering exploits.

Financial Institution Employees

- Privileged Roles
  - Administrative accounts
  - Business privileged users
  - Executives
- Non-Privileged Employees
  - Customer support

External Entities

- Financial institution consumers and clients
- Third-party employees
- General public
- Public figures

External entities are targets of social engineering campaigns too, but they may not be as well-protected or well-trained as financial services firm employees to recognize deepfake threats, making external individuals more exposed to social engineering tactics.

Financial institutions should identify their likely recipients of deepfake attacks. This knowledge may be applied to appropriate training and awareness programs and help develop rapid communication strategies to maintain trust and mitigate the impact of such attacks.

Deepfake Taxonomy: Threats to Institutions and Technologies

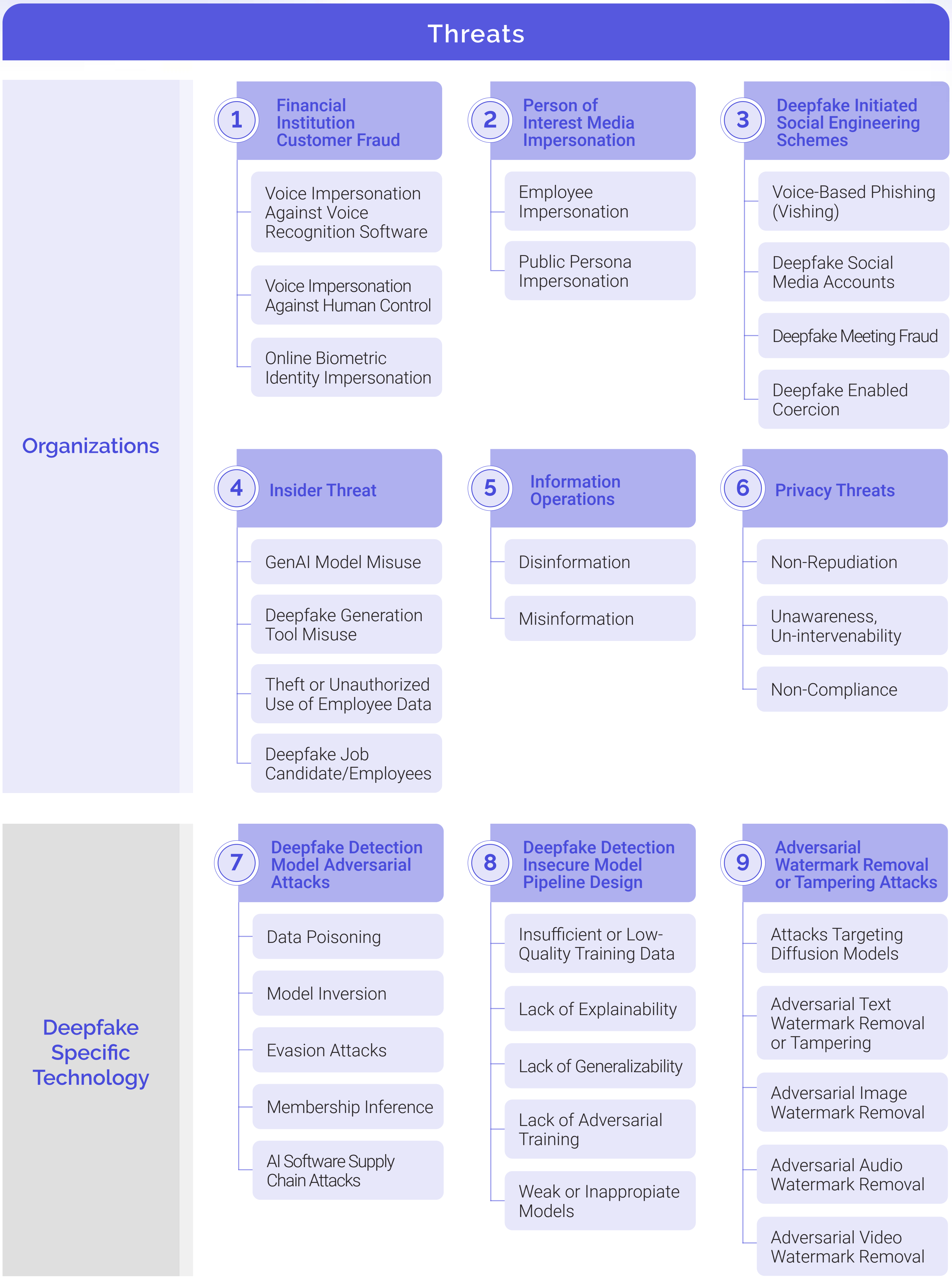
The first six categories of the FS-ISAC Deepfake Threat Taxonomy (presented below) focus on threats to financial institutions from deepfakes. The last three categories regard technologies specific to deepfakes. All the categories have a common trait: the abuse of trust.

The sector runs on trust, and so do business operations. Employees must have trust in biometric authentication methods, the customer on the phone, the colleagues in the video conference, and the expression of consumer sentiment on social media to perform their job duties effectively.

Note that in each of the deepfake attack examples presented in the table above, the deepfake was used to engender greater trust to help the attacker achieve their goal.

> Abuse of trust is not a novel attack against the sector, but deepfakes leverage and exploit trust at a new level.

Those examples of deepfake attacks, their probable victims, and the financial institution assets likeliest to be threatened help to contextualize the FS-ISAC Deepfake Threat Taxonomy below. This Taxonomy categorizes the different types of attacks so that financial institutions can determine what they consider to be of greatest risk and therefore which controls they should put in place.



Threats

Organizations

1. **Financial Institution Customer Fraud**
   - Voice Impersonation Against Voice Recognition Software
   - Voice Impersonation Against Human Control
   - Online Biometric Identity Impersonation

2. **Person of Interest Media Impersonation**
   - Employee Impersonation
   - Public Persona Impersonation

3. **Deepfake Initiated Social Engineering Schemes**
   - Voice-Based Phishing (Vishing)
   - Deepfake Social Media Accounts
   - Deepfake Meeting Fraud
   - Deepfake Enabled Coercion

4. **Insider Threat**
   - GenAI Model Misuse
   - Deepfake Generation Tool Misuse
   - Theft or Unauthorized Use of Employee Data
   - Deepfake Job Candidate/Employees

5. **Information Operations**
   - Disinformation
   - Misinformation

6. **Privacy Threats**
   - Non-Repudiation
   - Unawareness, Un-intervenability
   - Non-Compliance

Deepfake Specific Technology

7. **Deepfake Detection Model Adversarial Attacks**
   - Data Poisoning
   - Model Inversion
   - Evasion Attacks
   - Membership Inference
   - AI Software Supply Chain Attacks

8. **Deepfake Detection Insecure Model Pipeline Design**
   - Insufficient or Low-Quality Training Data
   - Lack of Explainability
   - Lack of Generalizability
   - Lack of Adversarial Training
   - Weak or Inappropriate Models

9. **Adversarial Watermark Removal or Tampering Attacks**
   - Attacks Targeting Diffusion Models
   - Adversarial Text Watermark Removal or Tampering
   - Adversarial Image Watermark Removal
   - Adversarial Audio Watermark Removal
   - Adversarial Video Watermark Removal

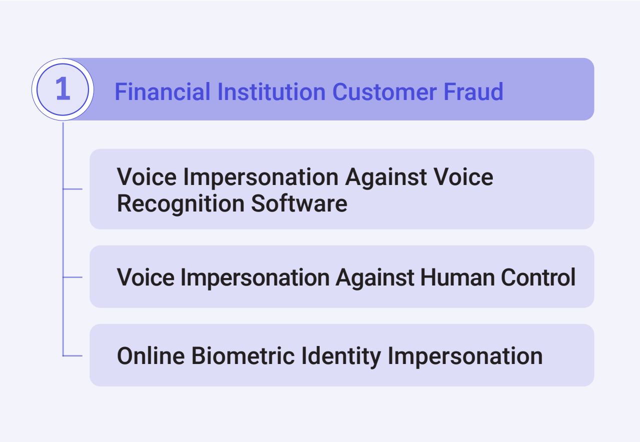
Figure 2. FS-ISAC Deepfake Threat Taxonomy.

The FS-ISAC Deepfake Taxonomy covers two topics:

- The six threats that financial services firms face from deepfakes.
- Three primary attack vectors targeting the technologies that detect and prevent deepfakes.

We describe each of these threats in more detail below and provide a summary of possible controls and mitigations in the following section.

1 **Financial Institution Customer Fraud:** One of the primary risks of deepfakes to financial services institutions involves **voice impersonation against voice recognition software,** in which attackers use synthetic voices to trick systems designed to verify identity through voice. Similarly, deepfake technology can be used to mimic human voices to bypass human controls and deceive individuals directly. That allows attackers to commit fraud by, for instance, impersonating customers who have dedicated contacts and wealth managers. Our Taxonomy labels such deception as "**voice impersonation against human control.**" Another serious threat is **online biometric identity impersonation,** in which attackers replicate biometric markers like facial features or fingerprints to gain unauthorized access.

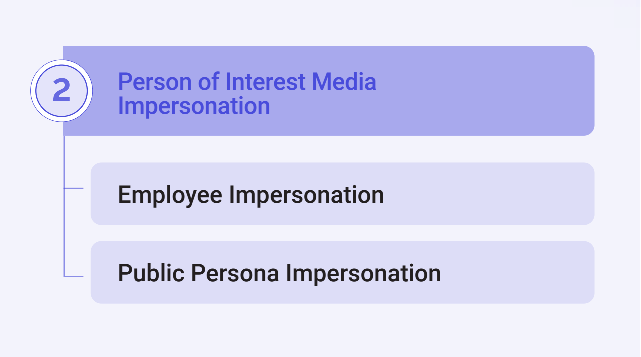

1 Financial Institution Customer Fraud

Voice Impersonation Against Voice Recognition Software

Voice Impersonation Against Human Control

Online Biometric Identity Impersonation

2 **Person of Interest Media Impersonation:** Key corporate officials are at risk of **employee impersonation,** in which an attacker uses deepfake technology to convincingly mimic an employee's appearance or voice to breach security. In **public persona impersonation,** the impersonated individual is usually a celebrity, corporate leader, or public official and the malicious purpose is to spread misinformation, damage public perception of the victim, or harm the institution's reputation.

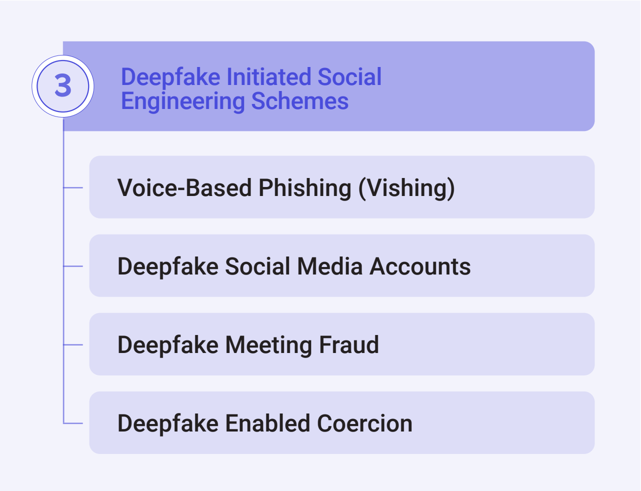

2 Person of Interest Media Impersonation

Employee Impersonation

Public Persona Impersonation

3 **Deepfake Initiated Social Engineering Schemes:** Deepfakes enhance the effectiveness of social engineering schemes significantly and can be deployed in various attacks. **Voice-based phishing (vishing)** social engineering attacks are phone calls – often highly targeted – meant to cause the victim to perform some action of benefit to the threat actor. Vishing is particularly convincing when it includes the deepfaked voice of someone familiar to the victim.

Deepfake social media accounts appear to be legitimate, tricking individuals into divulging sensitive information. **Deepfake meeting fraud** inserts synthetic participants in virtual meetings, potentially adding legitimacy to a targeted attack. **Deepfake enabled coercion** employs manipulated media to blackmail or extort employees into performing actions that help a perpetrator. Consumers, too, can be victims of social engineering schemes that leverage deepfakes, such as impersonations of a victim's family member claiming to be in trouble and in urgent need of money.


3 Deepfake Initiated Social Engineering Schemes

Voice-Based Phishing (Vishing)

Deepfake Social Media Accounts

Deepfake Meeting Fraud

Deepfake Enabled Coercion

4 **Insider Threat:** Employees can **misuse GenAI models** or **misuse deepfake generation tools** to perform an attack that leverages their knowledge of the organization and access to assets. **Theft or unauthorized use of employee data** allows employees to create a deepfake externally. Similarly, threat actors can create **deepfake job candidates and/or employees** to hide their identity and obtain access to sensitive information or restricted corporate systems.

4 Insider Threat

- GenAI Model Misuse
- Deepfake Generation Tool Misuse
- Theft or Unauthorized Use of Employee Data
- Deepfake Job Candidate/Employees

5 **Information Operations:** Deepfakes are effective tools of **disinformation** and **misinformation**. Disinformation is meant to deceive or mislead via the intentional dissemination of false information. Misinformation involves spreading false information without intent to deceive. Both can damage public perception of the subject, trust in institutions, and the overall integrity of information.

5 Information Operations

- Disinformation
- Misinformation

6 **Privacy Threats:** Deepfakes threaten privacy in a number of ways, including **non-repudiation** issues in which individuals cannot deny the authenticity of manipulated media attributed to them, **unawareness** in which individuals don't know that deepfakes of them have been created or are being used against them, or the reverse, **un-intervenability**, in which they know of the deepfake attack but lack the means to recover it. **Non-compliance** with regulations and privacy standards is another threat, in the sense that non-compliance removes obstacles to deepfake attacks.

6 Privacy Threats

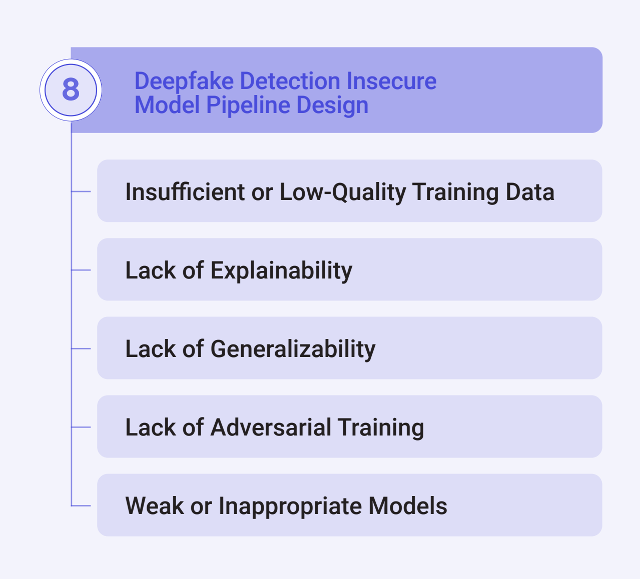
- Non-Repudiation
- Unawareness, Un-intervenability
- Non-Compliance

7 **Deepfake Detection Model Adversarial Attacks:** This category represents threats against models intended to detect deepfakes (e.g., audio, video). It includes **data poisoning**, in which the training data is tampered with to mislead the model. **Model inversion** extracts the data from the detection model with the goal of reproducing a replica of the model. **Membership inference** is similar but intends to learn information about specific training data instances. **Evasion attacks** occur when adversaries input specially crafted data to deceive (or evade) the detection system. AI software supply chain attacks refer to the exploitation of supply chain vulnerabilities to compromise the model, such as compromising widely used software libraries.

7 Deepfake Detection Model Adversarial Attacks

- Data Poisoning
- Model Inversion
- Evasion Attacks
- Membership Inference
- AI Software Supply Chain Attacks

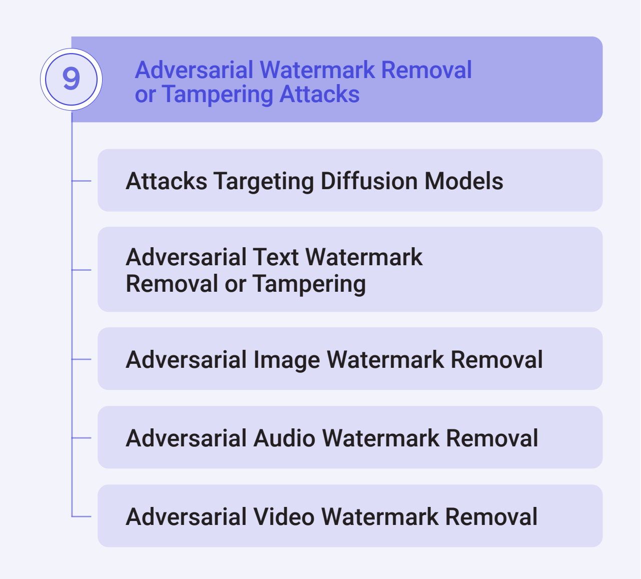
8 **Deepfake Detection Insecure Model Pipeline Design:** Flaws in detection model design and implementation can allow adversaries to bypass deepfake detection models. For example, **insufficient training data** and a **lack of high-quality training data** can reduce the model's overall performance and effectiveness. **Lack of explainability** makes understanding and interpreting the model's decisions difficult (which can also be the result of a substandard model pipeline), whereas **lack of generalizability** inhibits the model's performance across different scenarios and datasets. **Lack of adversarial training** can allow bypass or evasion attacks where the attacker can avoid detection by the model. **Weak or inappropriate models** refers to the absence of combined models that could improve robustness or the use of models that are inappropriate for the detection task. This also includes the weak model performance that can arise from a model pipeline that does not evaluate performance in a manner that gives a clear idea of how well the model performs. To learn more about risks associated with generative AI models, please see the FS-ISAC AI Risk Working Group's white papers, [Financial Services and AI: Leveraging the Advantages, Managing the Risks](#).


8 Deepfake Detection Insecure Model Pipeline Design

- Insufficient or Low-Quality Training Data
- Lack of Explainability
- Lack of Generalizability
- Lack of Adversarial Training
- Weak or Inappropriate Models

9 **Adversarial Image Watermark Removal or Tampering Attacks** affect a mitigation strategy applied to preventing the creation of deepfakes. Watermarks are modifications to images, audio, video, and text that are imperceptible to humans but that can impact computer interpretation of media. For example, watermarks can be used to prove the provenance of a piece of media, or they can be used to modify the media in such a way that it is difficult to generate high-quality deepfakes from them. Attacks that attempt to modify or remove the watermarks can therefore lead to media that can be used in the creation of deepfakes.

Attacks targeting diffusion models manipulate individual images either before or after their creation to remove or alter watermarks. **Adversarial text watermark removal or tampering attacks** compromise the integrity of watermarks – i.e., certain words used at predetermined frequencies – inserted into the output of large language models. **Adversarial audio watermark removal** obscures or distorts the watermark signal, or selectively removes samples to degrade the watermark's integrity. Similarly, common video processing operations and transformations, such as cropping or compression, can be used for **adversarial video (or image) watermark removal.**


9 Adversarial Watermark Removal or Tampering Attacks

- Attacks Targeting Diffusion Models
- Adversarial Text Watermark Removal or Tampering
- Adversarial Image Watermark Removal
- Adversarial Audio Watermark Removal
- Adversarial Video Watermark Removal

Summary of Controls to Protect Against Deepfake Attacks

Understanding the different types of threats posed by deepfakes and how they can be taxonomized clarifies the types of controls most suitable to defense. Below, we summarize controls identified in Figure 3.

Controls are classed as either prevention or detection. Controls in green represent prevention controls. Controls in blue represent detection controls. Both types of controls are important: prevention controls help, but they can't prevent all deepfakes, so detection controls are necessary to identify active exploits.

Financial services institutions should perform a complete threat modeling for each of the threat categories. Some attack vectors can be sufficiently protected such that deepfake detection is not required.

> Some consumers – such as those with speech disorders – use Augmentative Alternative Communication tools. Deepfake detection controls unable to distinguish between those tools and illegitimate uses may flag these technologies as adversarial. Legitimate consumers may find such mistakes frustrating and insulting, so additional authentication controls are recommended to ensure these tools are covered appropriately.

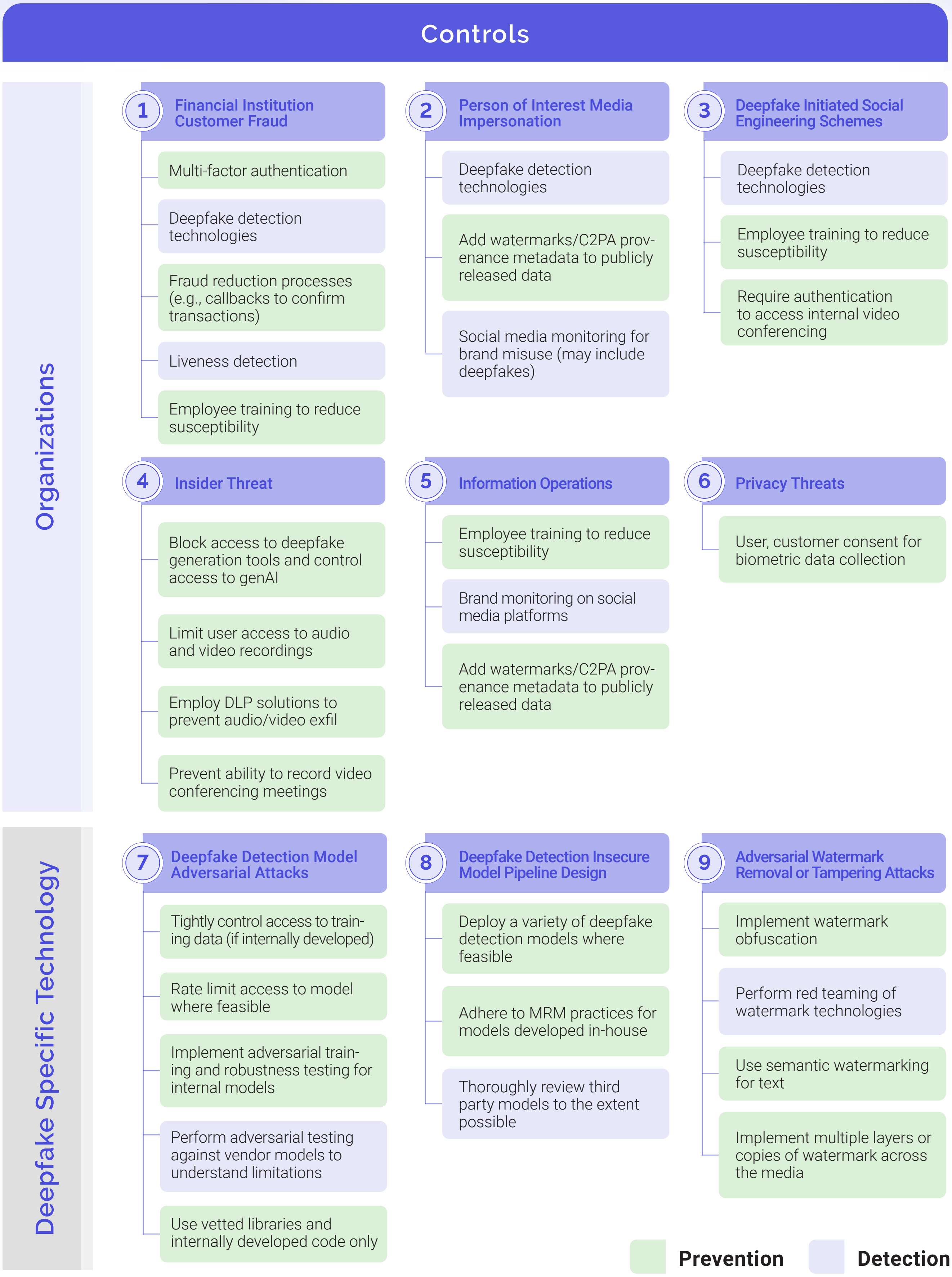
Some controls protect against multiple types of attacks – most notably, employee training. Employees should understand what deepfakes are, the basics of recognizing them, and how to report them. Contextualizing deepfake training is likely to increase employees' vigilance – no one should assume they won't be targeted – so align training to the role with examples of deepfakes the employee is most likely to receive.

Watermarks are a new protection capability that appear in multiple categories; however, this is still very much a nascent control and the Taxonomy lists threats to this protection capability (category 9). We expect that watermark controls and threats will continue to evolve considerably.

Authentication is a very important control. It appears in the first category as "multi-factor authentication" applying to customer identity confirmation (rather than permitting only one form of identification, such as the customer's voice). In the third category, authentication controls regard attendees of web conferences. Having knowledge of the meeting isn't sufficient authentication, nor is looking familiar to colleagues.

Examples of Role-Based Deepfake Training:

- Wealth managers should learn to be alert to deepfakes of high-net-worth clients asking for a funds transfer.
- Employees in accounting functions should watch for deepfakes of executives requesting a confidential funds transfer.



Controls

Organizations

1 Financial Institution Customer Fraud
- 🟩 Multi-factor authentication
- 🟪 Deepfake detection technologies
- 🟩 Fraud reduction processes (e.g., callbacks to confirm transactions)
- 🟪 Liveness detection
- 🟩 Employee training to reduce susceptibility

2 Person of Interest Media Impersonation
- 🟪 Deepfake detection technologies
- 🟩 Add watermarks/C2PA provenance metadata to publicly released data
- 🟪 Social media monitoring for brand misuse (may include deepfakes)

3 Deepfake Initiated Social Engineering Schemes
- 🟪 Deepfake detection technologies
- 🟩 Employee training to reduce susceptibility
- 🟩 Require authentication to access internal video conferencing

4 Insider Threat
- 🟩 Block access to deepfake generation tools and control access to genAI
- 🟩 Limit user access to audio and video recordings
- 🟩 Employ DLP solutions to prevent audio/video exfil
- 🟩 Prevent ability to record video conferencing meetings

5 Information Operations
- 🟩 Employee training to reduce susceptibility
- 🟪 Brand monitoring on social media platforms
- 🟩 Add watermarks/C2PA provenance metadata to publicly released data

6 Privacy Threats
- 🟩 User, customer consent for biometric data collection

Deepfake Specific Technology

7 Deepfake Detection Model Adversarial Attacks
- 🟩 Tightly control access to training data (if internally developed)
- 🟩 Rate limit access to model where feasible
- 🟩 Implement adversarial training and robustness testing for internal models
- 🟪 Perform adversarial testing against vendor models to understand limitations
- 🟩 Use vetted libraries and internally developed code only

8 Deepfake Detection Insecure Model Pipeline Design
- 🟩 Deploy a variety of deepfake detection models where feasible
- 🟩 Adhere to MRM practices for models developed in-house
- 🟪 Thoroughly review third party models to the extent possible

9 Adversarial Watermark Removal or Tampering Attacks
- 🟩 Implement watermark obfuscation
- 🟪 Perform red teaming of watermark technologies
- 🟩 Use semantic watermarking for text
- 🟩 Implement multiple layers or copies of watermark across the media

🟩 **Prevention** 🟪 **Detection**

Figure 3. Summary of Controls Against Deepfake Attacks.

Conclusion

The rapid advancement of deepfake technology presents a complex and evolving landscape of threats to financial institutions and their stakeholders. Our Taxonomy illustrates the multifaceted nature of these challenges, spanning customer authentication vulnerabilities to executive impersonation risks and sophisticated information operations.

At the core of these vulnerabilities lie weaknesses in authentication mechanisms, data verification processes, and user awareness. These gaps create fertile ground for deepfake exploitation, underscoring the need for a comprehensive approach to security that addresses technical, procedural, and human factors alike. Diverse threat vectors require diverse mitigation strategies, tailored to address each concern.

While the financial sector has begun to develop and implement various security controls and mitigation strategies, many of these countermeasures remain in their early stages. This immaturity leaves a considerable residual risk that must be actively managed. The dynamic nature of the threat landscape demands ongoing vigilance, research, and adaptation of layered security measures to keep pace with evolving attack methodologies.

The role of education and awareness in combating deepfake threats cannot be overstated. By fostering a culture of vigilance and critical thinking, financial institutions can create a human firewall that complements technological defenses. This approach is particularly crucial given the sophisticated and often persuasive nature of deepfake social engineering content.

As financial institutions increasingly rely on biometric authentication methods and digital communications, the risk of deepfake-enabled fraud escalates. Users of voice recognition systems, once considered a secure form of authentication, now face the challenge of distinguishing between genuine customer voices and highly convincing synthetic replicas. Similarly, video conferencing and digital messaging platforms, crucial for modern finance operations, become potential avenues for sophisticated impersonation attacks.

Addressing the deepfake challenge is not a task that any single institution can accomplish in isolation. It requires a collaborative effort involving financial institutions, technology providers, regulatory bodies, and industry groups. By sharing knowledge, developing standards, and coordinating responses, the financial sector can present a united front against this emerging threat.

In conclusion, while the threat posed by deepfakes to financial institutions is significant and evolving, a proactive, multi-faceted approach to security can substantially mitigate these risks. The path forward lies in the continuous improvement of detection technologies, coupled with robust security practices and comprehensive awareness programs.

We encourage you to share your own deepfake knowledge, best practices, and experiences for the benefit of the sector. By sharing information and maintaining vigilance in the face of this dynamic threat, the financial sector can preserve trust, protect assets, and ensure the integrity of its operations in an increasingly digital world.

Contributors

Hiranmayi Palanki, American Express

Dr. John T. Hancock, American Express

Benjamin Dynkin, Wells Fargo

Cassi Hutchinson, Fidelity

Elizabeth Geary, Fiserv

Lisa Matthews, Ally

Monica Maher, Goldman Sachs

Marc Teenie, LeTort Management & Trust

Dr. Carrie E. Gates, FS-ISAC

Mike Silverman, FS-ISAC

Endnotes

[1] Executive survey: https://www.business.com/articles/deepfake-threats-study/

[2] Deepfake fraud losses: https://www2.deloitte.com/us/en/insights/industry/financial-services/financial-services-industry-predictions/2024/deep-fake-banking-fraud-risk-on-the-rise.html