

DG – DeepfakesIDV Discussion Group

**Proposed Work Product**

Maxine Most, October 22, 2024

## Deepfake Detection, Protection and Countermeasures for Remote Identity Verification (PowerPoint Document)

### Executive Summary (Denny)

### Introduction

## **Context**

* + - Regulatory Environment – link to Appendix: Regulations
		- Standards - link to Appendix: Relevant Standards
* Audience – Identify key constituencies:
* Enterprise: IT, Security, Cybersecurity
* Vendors: Biometric, IAM, CIAM, Fraud, Cyber3eecurity,
* Policy Makers: Government and NGO’s in Privacy, Cybersecurity,
* Standards/testing/certifications bodies

### Content

## **RIDV Process**

RIDV Process Diagram

* Define and explain of each step
* Relevant technologies and solutions (Generic)

**Introduction**

Realistic simulations of human voice, video and text created by “Generative AI” systems have proliferated over the last year. The simulations pose enhanced risk that Identity Proofing and Verification (IDPV) systems will be unable to distinguish real from fake signals. Organizations that rely on IDPV services to prevent fraud or impersonation are experiencing higher number and frequency of fraudulent attempts.

The Kantara Deepfake-IDV Discussion Group—*Deepfake Threats to Identity Verification & Proofing*—was formed in September 2023 to explore how IDPV (Identity Proofing and Verification) systems could be subverted or fooled by “deepfakes,” “Generative AI,” and other AI-related mechanisms.

The group primarily comprised technical experts from within the biometric and digital identity marketplace, including vendors, individual subject matter experts, and contributors from end-user organizations.

The anticipated output of the discussion group was a report describing the nature of the threats, vulnerabilities, and potential countermeasures designed to:

* Inform various non-technical stakeholders, Govt. & Enterprise policymakers, and purchasers of IDPV services about AI-related Deepfake attack techniques that may disrupt their services
* Enable readers to discuss the topic and potential risk mitigation actions within their organization and with IDPV service providers.

This document represents the group’s output and reflects the group’s consensus thinking about deepfake detection, protection, and countermeasures for Remote Identity Verification (RIDV).

**Core Concepts**

**What is Artificial Intelligence (AI)?**

The Oxford Dictionary defines Artificial Intelligence as the theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.

AI has undergone significant evolution, starting from simple statistics to the complex artificial intelligence systems we see today. Initially focused on pattern recognition and prediction through statistical methods, AI has expanded to include data science, machine learning, and deep learning. Recent advancements in computing power, data availability, and research have accelerated AI development, leading to groundbreaking applications in various fields, particularly in identity verification (IDPV).

Two major innovations driving AI today are Transformer Models and Generative Adversarial Networks (GANs). Transformer Models, such as those used in natural language processing applications like ChatGPT, utilize neural networks to understand context and process vast datasets, enabling capabilities like conversation simulation and text summarization. GANs, consisting of a generator and a discriminator, create realistic content, including images, videos, and audio, which has significant implications for generating synthetic media and deepfakes.

AI technologies like computer vision, biometrics, and natural language processing are revolutionizing the Identity & Access Management (IAM) field. Computer vision techniques such as optical character recognition (OCR) and object detection enhance document verification, while biometrics, including face, fingerprint, and voice recognition, provide robust identification methods. Additionally, pattern and anomaly detection, behavioral analysis, and risk scoring improve security and fraud detection. As AI continues to evolve, its utility in both ethical and unethical identity-related applications has grown significantly. it is crucial to adapt and innovate to address emerging challenges and leverage new opportunities in IDPV.

Artificial Intelligence, while researched for a long time, has only recently emerged as capable or viable enough to present a significant operational threat. Reacting to this recent emergence, relevant standards, regulation and legal controls are nascent, developmental and unstable.

**What is an “AI-Deepfake”?**

The term “Deepfake” is effectively industry jargon and is not universally technically defined. What might be described as a “Deepfake” usually reflects specific context that does not automatically reflect the entirety of potential contexts. Thus, it may be helpful to recognize a deepfake as the product of using AI to generate a realistic imitation of something legitimate, that may or may not be physically, factually or contextually accurate. As that, an “AI Deepfake” (deepfake) is inherently deceptive and, therefore, can be used covertly or overtly to achieve ethical or unethical goals.

**Deepfake Threats and Attack Vectors in the Context of Identity & Access Control**

**What is an Identity?**

The meaning of IDENTITY is the distinguishing combination of physical, biographical and personality characteristics of an individual human. Identity is the set of qualities, beliefs, personality traits, appearance, and/or expressions that characterize a person or a group. Identity encompasses the memories, experiences, relationships, and values that create one's sense of self. An **Identity Attribute** is a singular, specific distinguishing physical, biographical or personality characteristic of an individual human. A **Legal Identity** is a standardized combination of specific Identity Attributes that a governmental Identity Issuing Authority uses to identify a unique individual within its jurisdiction. A Legal Identity represents the highest confidence identity data available to describe an identity, because it has Integrity, is relatively immutable and is long lived. Legal Identities provide the highest level of trust the identity is real.

**What is Identity Verification (IDV)?**

Within the Identity & Access Management (IAM) context, **Identification** is the process of establishing an individual’s uniqueness within the human population. **Identity Verification (IDV)** systems are systems that attempt to prove the legitimacy of an identity, and its relationship to an individual physical human within that population. In other words, IDV systems attempt to determine if a person (**a Claimant**) is who they claim to be. They do this my determining the validity of presented Identity Attributes and the validity of their relationship to the Claimant. **Authentication** is the process of attempting to ensure the person attempting to access granted privileges is the same person that was granted the privileges. Ideally, authentication mechanisms are best selected after a satisfactory identity verification has been completed. Further, stakeholders should view Identification, Verification and Authentication as codependent. The ideal Authentication mechanism is derived from the Verification mechanisms used during the Identity Verification process. This allows Authentication to be based on multiple proven relationships between the verified identity, the verifier, authenticator and the Claimant.

There are **Supervised** and **Remote Unsupervised** Identity Verification Systems. Supervised IDV implies an in-person human inspection of a Claimant and presented Identity Attributes to verify the Claimants identity. Remote Unsupervised (RIDV) systems imply an automated inspection of a Claimant and presented Identity Attributes to verify the Claimants identity.

While there are mature technology, performance and conformance standards for some elements within an Identity Verification system, there are no such standards or specifications for a complete, end-to-end Identity Verification system.

**Where and how do AI-Deepfakes fit in Identity Verification Schemes?**

Within the IAM context, **Deepfake Attacks** are the creation of highly realistic but fake audio content, video content, and still images, that can be used to compromise IDV systems in various ways. Attackers may use deepfake technology to present falsified identities, modify genuine documents, or create synthetic personas, exploiting weaknesses in the verification process. Deepfake attacks are used to gain unauthorized access to privileges associated with a wide range of assets, services, systems, and domains, from social and traditional media to national security and human rights to banking and access to digital systems. Remote Identity Verification systems are particularly and increasingly at risk from the evolving threats of deepfake technology.

**Types of Deepfake Attacks**

Deepfakes are not “one thing,” but rather a class of AI-produced digital face, voice and document manipulation and / or synthesis which is of sufficient quality to fool both human and automated attribute validation systems to undermine a remote digital identity scheme.

There are two primary Deepfake attack vectors, **Presentation Attacks** and **Injection Attacks**. Presentation Attacks present deepfake and other fake identity data directly to a sensor, like holding a fraudulent photo to a camera, in lieu of your legitimate selfie photo. These leave the system integrity unaffected, but rely on relatively poor ability of the system to discern real and fake identity attributes. A Deepfake Presentation Attack would include presenting an AI-enhanced or generated image or video (super high-definition screen) directly to a sensor, like a camera, in lieu of a living human Claimant. Injection Attacks specifically involve a combination of hardware and software to violate or disrupt system integrity, to replace or manipulate legitimate data with fake data. Injection Attacks typically bypass and eliminate system sensors, like a camera, replacing it by introducing (injecting) fake identity attribute data, like a fake digital photo image or video, into the system in lieu of what the sensor would have collected. Common AI-Deepfake Injection techniques include:

* **Virtual Camera Injection**: Using software to create a virtual camera that directly feeds pre-recorded or synthetic video into the biometric system.
* **Device Emulation**: Emulating a mobile device to manipulate the system calls and inject fraudulent biometric data.
* **Function Hooking**: Altering the flow of system calls to replace legitimate biometric data with fraudulent data.
* **Man-in-the-Middle (MitM) Attacks**: Intercepting and altering the data stream between the biometric sensor and the processing system.

Presentation Attacks were the first “Spoof Attack” vector and still represent the vast majority of attacks. However, Injection Attacks are substantially more sophisticated and infinitely scalable. Importantly, Injection Attacks can be fully automated. Consequently, Deepfake Injection attacks are widely viewed as a significantly more dangerous threat, and are expected to surpass Presentation attacks as the most utilized attack relatively quickly.

This summary explores the broad types of deepfake attacks, specific methods such as face swaps, expression swaps, synthetic imagery, and synthetic audio, and the various points of attack, including physical presentation attacks, injection attacks, and insider threats. Understanding these threats is crucial for developing robust defenses against the manipulation of identity verification systems. *Here, we will focus specifically on deepfake threats and attack vectors in the scope of remote identity verification.*

1. **Still Imagery Deepfakes:**
	* **Face Swaps:** Replacing a person's face in a video with another person's face, often seamlessly.
	* **StyleGAN2-Type Synthetic Imagery:** Generating highly realistic human faces that do not belong to any real person.
	* **Diffusion-Based Imagery (e.g., Midjourney, Stable Diffusion):** Creating realistic images from textual descriptions or other input images, making it possible to fabricate convincing identity photos.
2. **Audio Deepfakes:**
	* **Synthetic Speech (e.g. Eleven Labs):** Creating realistic synthetic voices using tools to impersonate someone during a voice verification process.
	* **Voice Cloning:** Replicating someone's voice to bypass voice authentication systems.
3. **Video Deepfakes (which may include a combination of the still imagery and audio techniques discussed above)**
	* **Expression Swaps:** Altering the facial expressions of a person in a video to match those of another person.
	* **Next-Gen Video Avatars (e.g. Synthesia, HeyGen):** Creating fully synthetic avatars that can move and speak like real humans, making it hard to distinguish between real and fake identities.
4. **Document Deepfakes:**
	* AI Generated imitations of physical legal documents, including legitimate stolen identity data and/or synthetic illegitimate identity data, uploaded to supervised and unsupervised IDV systems to misrepresent the claimant as a legitimate and unique individual within the population.

**Points of Attack**

The following diagram summarizes key attack points for deepfakes in remote identity verification (IDV). It follows from earlier work on codifying possible presentation attack points in biometric systems as included in ISO/IEC 30107-1:2016 and other industry research, such as that done by Stephanie Schuckers at CITeR.

Open image-20240731-163323.png



**Evolution of Deepfake Attacks**

No doubt, while significant risks and attack vectors exist today, the threats are growing rapidly due to the remarkable pace of technology advancement, in particular with and around Generative AI.

* **Improved Visual and Audio Quality:** Deepfake technology is rapidly advancing, producing increasingly realistic and convincing fake content that is harder to detect.
* **Personalization and Targeting:** Deepfakes are becoming more sophisticated in mimicking specific individuals, making them more effective in targeted attacks.
* **Integration with Other Attack Vectors:** Deepfakes are being combined with other cyber-attack methods, creating more complex and difficult-to-detect threats.
* **Real-Time Manipulation:** Advancements in processing power and algorithms are enabling real-time deepfake creation and manipulation, posing new challenges for detection.
* **Scalability and Behavioral Mimicry:** Deepfake technology is evolving to replicate not just appearance and voice, but also mannerisms and behaviors, making detection even more challenging.

**Conclusion**

Deepfake threats present significant challenges to remote identity verification systems. The broad range of attack vectors, from audio and video to still imagery, and the specific methods employed, such as face swaps, expression swaps, synthetic imagery, and synthetic audio, underscore the sophistication and potential impact of these threats. Points of attack span physical presentation, data injection, and live interactions, highlighting the need for robust security measures and continuous advancements in detection technologies to mitigate the risks posed by deepfakes.

# **Prevention, Detection and Countermeasures**

RIDV Process Diagram with Attack Vectors Overlay + Preventions, Detection, Countermeasures Overlay

* Define and explain Prevention, Detection, and Countermeasures to applied at each step
* Detail relevant/related processes, technologies, and solutions (Generic)

[Prevention and Countermeasures](https://kantara.atlassian.net/wiki/x/IQCjE)

**Introduction**

In the previous section, we more narrowly described Deepfakes as the creation of highly realistic but fake audio, video, and still images that can be used to compromise IDV systems in various ways. Attackers may use deepfake technology to present falsified identities, modify genuine documents, or create synthetic documents or personas, exploiting weaknesses in the verification process. To follow, we described various potential IDV system weaknesses and respective attacks, enabled by those weaknesses. This summary attempts to describe the most widely accepted perspectives, policies, technologies and related methods to detect and defend against Deepfake attacks in Remote Identity Verification Systems.

**Think Holistically and Proactively**

There is no silver bullet to defend against AI-deepfakes. There is no single product or technology to effectively defend against all AI-deepfakes. Attacks will emerge in as many forms and vectors as human creativity, aided by advanced Artificial Intelligence can imagine and develop. Moreover, AI-deepfake attacks will evolve as rapidly as human creativity, aided by advanced Artificial Intelligence and support. Effectively, AI has accelerated and will continue to accelerate the perception of, development, effectiveness and deployment of new AI-Deepfake Injection attacks. And further yet, you don’t know what you don’t know. Reactionary responses could be effective against previously endured and studied attacks, damage be done. They likely cannot defend against new forms of attacks. It’s imperative to expect the unexpected. Defending against AI-deepfakes requires thinking and planning ahead; thinking proactively to mitigate the possibility of an attack, before it occurs.

Many thought leaders in the space recommend a more holistic approach to deepfake attacks, perceiving the threat from the perspective of “The Three P’s”; **Policy, Provenance and Protection.** Stakeholders should assume that the “State of the Market” is likely far behind the “State of the Art” in AI-Deepfake mitigation. Moreover, Stakeholders should assume solution vendors may not represent either State of efficacy. Further still, Stakeholders should assume solution providers may not understand their products efficacy or even misrepresent its efficacy. Therefore, any Stakeholders first priority should be to enact Policy to achieve and maintain a heightened state of awareness or even expertise regarding current threat vectors. In order to achieve a sufficient state of readiness, Stakeholders should enact Policy to annually budget adequate capital to invest in and maintain such expertise. Stakeholder cyber-security teams should be sufficiently aware of the current States of the Market and Art for AI-Deepfake mitigation defenses and should set a Policy goal to attempt to achieve State of the Art capabilities.

**Plan to use the most robust IDV capabilities available….do the research first.**

* Knowing that a person (the “claimant”) is who they claim to be (identity) is important when deciding whether the person should get access to entitled privileges (service or domain access) or assets.
	+ There are different “Identities” used in different environments and interactions.
		- Biological
		- Legal
		- Social
		- Other?
* Identity Verification Systems are used to determine an individual’s uniqueness within a group (of humans) and typically include four or five primary components
	+ A human being
	+ An Identity for that human being
	+ Privileges assigned or entitled to that identity
	+ Credentials that bind the human to the identity and to the privileges
	+ Authenticators to prove a previously verified identity (optional)
* ID Proofing and Verification (IDPV) and Authentication systems are used to check whether a person is who they claim to be, primarily for privilege access control negotiations.
* IDPV core processes include:
	+ Verifying the claimant is an actual human being; “A bot, Deepfake, or human?”
		- In person interaction, Biometric Liveness Check, etc.
	+ Verifying the legitimacy of the claimed identity; “Does this identity actually exist?”.
		- Document and attribute validation
			* Effectively a background check, etc.
	+ Verifying the living human claimant is the person described by the verified identity; “We know this is a human. We know a person named Jay Meier exists. But is this human actually Jay Meier?”
		- Biometric verification, either in person or automated, against verified identity data (like a validated Passport photo)
			* Human brains evolved to recognize faces and voices. Voices can’t be represented in physical credentials. So, faces are the de facto standard binder between a human, an identity and entitled privileges.
* IDPV systems work by gathering information, documents, identity attributes and other data, like biometric data (photos etc.), from the claimant to verify against trusted sources like government records, publicly available/social data sources, proprietary or commercial databases, others.
	+ Some data sources are more trustworthy than others.
		- “You are who the Government says you are.” Government is the original issuer and arbiter of who we are in the real world.
			* Birth Certificates, Driver Licenses, Passports, National IDs are the common identity documents. However, these can be faked.
				+ Therefore, the source data at the government issuing authorities (like a DMV or State Department or National ID Office) are the strongest and most trustworthy identity data available.
* IDPV systems may be automated or manual (or both), use human adjudicators, cameras and other sensors, software and AI, or document validation systems, like OCR, PKI and other cryptographic systems.
* Identity Authenticators are derived from IDPV Systems and are typically the binding mechanism in credentials
* The primary attack vectors for identity related crimes typically include:
	+ Submission or presentation of fraudulent/fake credentials
	+ Transference of legitimate credentials and/or identity authenticators.
		- IDPV systems can be vulnerable to tricking the system (human adjudicator or the other system components into believing that a fake thing (like a forged document) is real or convincing them that things match (like a passport photo and the claimant's face) when they do not.
* Some important defenses or countermeasures include:
	+ Using devices and data that are provably untampered
	+ Using a human supervisor to monitor the use of data collection devices
	+ Checking information and images against official data sources that are known to have strong quality management processes in place
	+ Using software, hardware, and other means to determine whether a claimant is an actual live human (when the IDPV system is interacting with the claimant using automated means or the IDPV system is relying on uncontrolled devices)
		- This is generally called "biometric liveness detection [the term originated from techniques used to detect whether a living human was being sampled, for example fingerprints]. Advanced biometric liveness detection methods include:
			* 3D Geometry-Based Detection: This method analyzes a face's three-dimensional depth and geometry to produce a 3D representation, making it harder for attackers to use 2D photos or videos to spoof the system.
			* Phoneme-Viseme Mismatch Detection: This technique detects inconsistencies between spoken phonemes and corresponding mouth shapes, which are common in deepfake videos
	+ Mitigation includes several techniques:
		- Session Metadata Analysis: This involves analyzing metadata like GPS, accelerometer data, and network information to detect anomalies indicative of injection attacks.
		- Automated Artifact Detection: Using convolutional neural networks (CNNs) to identify video and audio data inconsistencies that suggest tampering.
		- Camera Anti-Tampering Measures: Implementing cryptographic methods to ensure the authenticity of the data captured by the camera.
		- Presentation Attack Detection (PAD) refers to the automated process used by biometric systems to detect and prevent presentation attacks, commonly known as "spoofs." These attacks involve using fake biometric samples, such as photographs, masks, or synthetic fingerprints, to impersonate someone else and gain unauthorized access to systems or data. PAD systems utilize hardware and software technologies to determine whether the presented biometric is genuine. A critical subset of PAD is liveness detection, which involves analyzing anatomical characteristics or involuntary and voluntary reactions to confirm that the biometric sample is being captured from a living subject present at the point of capture.
		- There are two liveness methods:
			* **Active**: Requires user participation through actions such as nodding, blinking, or turning their head.
			* **Passive**: Relies on involuntary human cues, like pupil dilation, without requiring any specific actions from the user.

**Require Independent IDV System and Component Performance and Conformance Validation**

**Ensure System and Sensor Integrity**

**Automated Biometric Liveness Detection and Matching**

This combined document summary emphasizes the importance of robust identity verification systems to prevent fraud, improve security, and enhance trust in both digital and physical environments. It highlights the critical need for advanced biometric security measures and anti-fraud strategies, particularly focusing on liveness detection and the binding of biometric data to verified accounts. The summaries discuss the current limitations and vulnerabilities in existing identity verification systems, advocating for improvements to enhance overall security and reliability.

One core concept is the emphasis on biometric liveness detection. The articles criticize current guidelines and suggest implementing server-side liveness checks and concurrent data collection to prevent spoofing and ensure the authenticity of biometric data. Another key idea is the role of anti-fraud and authentication measures in protecting enterprises, including best practices for eliminating knowledge-based authentication, integrating risk analytics, and developing comprehensive monitoring dashboards.

Lastly, the summaries highlight the vulnerabilities of relying solely on Public Key Infrastructure (PKI) for digital security. The PKI Fallacy illustrates how weak user authentication can compromise even the strongest cryptographic device authentication systems, advocating for enhanced identity verification methods, including biometric authentication, to address these vulnerabilities and improve overall security. Additionally, the articles touch upon philosophical considerations in defining human identity, emphasizing the complexity of identifying and classifying individuals in various contexts.

**Identity Data, Document and Credential Validation**

**OCR**

**Cryptography, mDL, VC and Humanity Tokens**

**Govt**

**Environmental Data Signal Analysis**

The document "Protecting Against Deep Fakes in the Public Switched Telephone Network" focuses on identifying and mitigating the risks posed by deep fake technology within telecommunication systems. It discusses various threat vectors including identity theft, account takeover, wire fraud, and healthcare data breaches, emphasizing the necessity of safeguarding sensitive personal and financial information. The document also highlights the potential financial, reputational, and national security impacts of deep fakes, underscoring the urgency for robust defense mechanisms.

To address these threats, the document proposes both in-band and out-of-band detection solutions. In-band detection involves real-time access to voice media and employs voice analytics technologies from vendors like Pindrop, Validsoft, and Microsoft (Nuance). Out-of-band detection includes methods such as remote proofing via NIST 800-63A standards, Lidar technology, video camera verification, fingerprint recognition, and geolocation tracking. These solutions aim to detect and prevent deep fake incidents at various capture points, such as Interactive Voice Response (IVR) systems and media relay control points.

### **TABLE: RIDV Process, Prevention, Detection and Countermeasures**

 Create Table that captures the information from the Conant sections above

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Step Name | Process | Key Technologies | Attack Vectors | Prevention Process/Tech | Detection Process/Tech | Countermeasures Process/Tech | Relevant Regulations | Relevant Standards |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |

### Key Findings & Recommendations

* Overall
* For each type of audience
* Enterprises
* Vendors
* Policy Makers
* Standards/testing/certifications bodies

### Appendices

* Glossary
* Regulations
* Technical References
* Standards