Deepfake-IDV Discussion Group

Deepfake Detection, Protection, and Countermeasures for Remote Identity Verification (RIDV)

November 2024

kantara INITIATIVE

Deepfake-IDV Discussion Group

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 purchasers of IDPV services about AI-related techniques that may decrease their effectiveness and
- 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).

Executive Summary

- **TO BE ADDED: 20,000 - foot view** of what we did, found, why it matters

Content Overview

Introduction and Scope

- Defining the Problem: Why deepfakes are a threat to identity verification
- Project Scope: Focusing on remote identity verification (RIDV)

RIDV Analysis

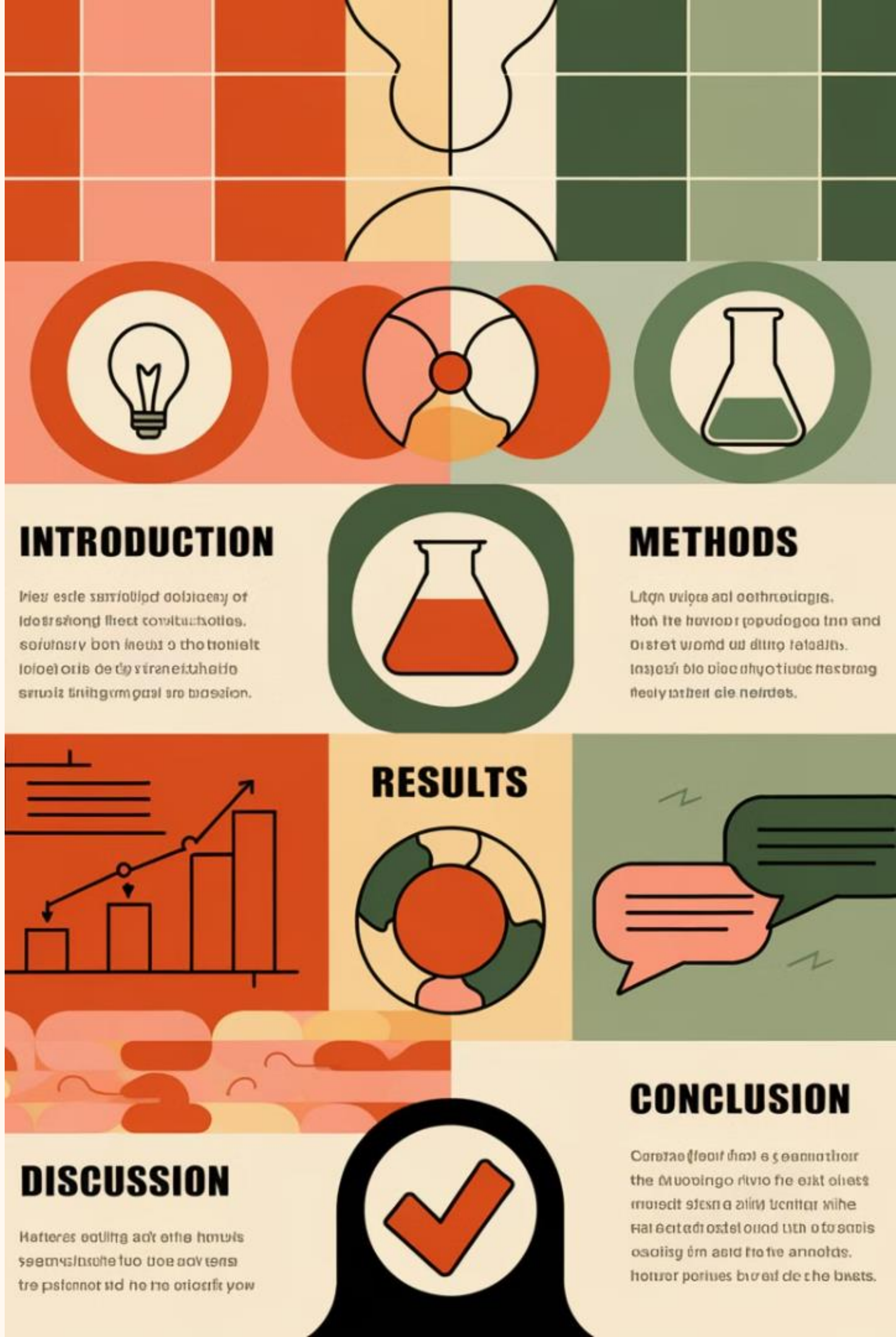
- Analyzing RIDV Processes: Understanding how RIDV works in practice
- Examining Attack Vectors: Identifying potential ways deepfakes can subvert RIDV systems
- Exploring Countermeasures: Proposing solutions to protect against deepfake attacks

Comprehensive Table

A table summarizing the RIDV process, attack vectors, and countermeasures

Conclusions

- Summarizing Key Findings and Recommendations: Presenting the group's key takeaways
- Appendices: Providing supplementary information and resources



Introduction

Introduction

Motivation for this Kantara Workgroup

The Kantara Deepfake-IDV Discussion Group was formed to explore how IDPV (Identity Proofing and Verification) systems could be subverted or fooled by "deepfakes," "Generative AI," and other AI-related mechanisms.

Key Terminology

This report uses "Deepfake" as a general term for any use of AI-based image manipulation to create a visual or audio representation of a person that does not accurately represent their real-world appearance or voice.

Introduction to Deepfakes

Deepfakes are created using Generative Adversarial Networks (GANs), a type of AI that trains on large datasets of images and videos. The AI learns to create new, synthetic data that is indistinguishable from real data. Deepfakes can be used to create realistic, but fake, images and videos of people. They can be used for a variety of purposes, including entertainment, political manipulation, and fraud.

Deepfakes Applications

Applications of deepfakes include creating convincing "fake news" videos, impersonating people in social media posts, and even creating fraudulent identity documents.

Benefits and Threats: Deepfake Spectrum

While deepfakes have potential benefits, such as for entertainment and education, they also pose significant threats to individuals and society. For example, they can be used to spread misinformation, damage reputations, and commit fraud.

State of IDV Market re: Deepfake Attacks

The IDV market is still developing its response to the growing threat of deepfake attacks. However, there are a number of promising countermeasures that are being developed, such as using biometrics to detect deepfakes, using AI to detect deepfakes, and using legal and ethical frameworks to regulate the use of deepfakes.

Audience for this Report

This report is intended for purchasers of IDPV services, as well as other stakeholders in the IDV ecosystem, such as policymakers, researchers, and technology developers. The report provides information about the threats posed by deepfakes to IDV systems, as well as potential countermeasures that can be used to mitigate these threats.

Motivation for this Kantara Workgroup

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.

This group will research how IDPV systems could be subverted or fooled by "deepfakes", "Generative AI", and other AI-related mechanisms.

The anticipated output of the discussion group is a report describing the nature of the threats, vulnerabilities, and potential countermeasures.

The report is intended to inform purchasers of IDPV services about AI-related techniques that may decrease the effectiveness of IDPV services, and to enable readers to discuss the topic and potential risk mitigation actions within their organization and with IDPV service providers.

Key Terminology: Identity

Identity

Distinguishing combination of physical, biographical, and personality characteristics of an individual human. It 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.

Identity Attribute

A **Identity Attribute** is a singular, specific distinguishing physical, biographical or personality characteristic of an individual human.

Legal Identity

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.

For a comprehensive list of Terminology see Appendix A: Glossary

Key Terminology: IDV

Identification

In the Identity & Access Management (IAM) context, **Identification** is the process of establishing an individual's uniqueness within the human population.

Identity Verification (IDV)

Identity Verification (IDV) systems are computer-based systems designed to prove the legitimacy of an identity and its relationship to an individual physical human within that population.

Remote Unsupervised Identity Verification (RIDV)

Remote Unsupervised (RIDV) systems imply an automated inspection of a Claimant and presented Identity Attributes to verify the Claimants identity.

Supervised IDV

Supervised IDV implies an in-person human inspection of a Claimant and presented Identity Attributes to verify the Claimants identity.

Authentication:

The process of attempting to ensure the person attempting to access granted privileges is the same person that was granted the privileges.

For a comprehensive list of Terminology see Appendix A: Glossary

Key Terminology: Deepfake

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

For a comprehensive list of Terminology see Appendix A: Glossary

Application of Deepfakes

Healthcare Training

Immersive Learning

Universal Translation

Personal Entertainment

Deepfake technology is not intrinsically problematic.

There are both legitimate and potentially beneficial applications of deepfake technology as well as inherently fraudulent, threatening and dangerous applications of of deepfake technology.

Financial Fraud

Election Interference

Cyberwarfare

The Deepfake Spectrum on the following page illustrates the various applications of deepfake technology