AU10TIX
IDENTITY INTELLIGENCE

| Whitepaper |

The Deepfake Wars

Smart defenders strike back

WWW.AU10TIX.COM

Contents

Foreword

By the time you read this, Generative-AI (deepfake) based identity fraud attacks will inevitably show new levels of sophistication.

We know it because we see it. How can one see what's increasingly invisible to both humans and machines?

Here's a paradigm shift: While intuition leads to developing AI tools to detect AI fakes (Kinda "anti-missile missile"), smart defenses add additional layers of detection that expose attacker scheme weak points (Kinda "drones against launchers...").

In the age where easily cloned deepfake humanoids are ALREADY used commercially and politically, multi-layered defenses are no less than imperative.

This whitepaper gives you an idea how.

Ofer Friedman
Chief Business Development Officer
AU10TIX

What if you can't trust your own eyes and ears?

Deepfakes aren't going anywhere. They're only going to become more sophisticated as technology advances and more common because it is easy for anyone to get their hands on deepfake tools. And, yes, not all deepfakes are designed with malicious intent. Some make for great entertainment. Think of Tom Hanks in Forrest Gump, artfully inserted into historical footage with JFK. Or take the world's number one Tom Cruise impersonator, who broke the internet with a series of viral videos. One shows 'Cruise' announcing his run for president in 2020 while literally running through a desert.

But entertainment is just that. Threats, on the other hand, are serious. Deepfakes, while potentially identifiable by a trained eye, are not easily distinguished by the untrained eye. Jordan Peele proved just how damaging unchecked content can be with his deepfake of President Obama back in 2018.

Fakes have been on the radar of public and private organizations for a while. Their rapid rise in prevalence has sparked conversations about how to handle them. Talks have led to several discoveries and ways of combating deepfakes, too, but there's a clear need for alignment. Greater cooperation between identity **verification** vendors and regulators must take hold to stop bad actors in their tracks.

While growing awareness about the harm deepfakes is a win, fakes have also presented the world with a tough question to answer. What happens when you can no longer trust your own eyes or ears?

Leveling the Playing Field

As sophisticated as deepfakes can be, there is a way to peel back the layers they adorn and stop them in their tracks. But to do so, you'll need to know what deepfakes are made of and why their twin characteristics make them so effective.

This white paper peels back those layers, exploring the anatomy of deepfakes. It also shows how detection methods fare against them. But there is a twist to hunting deepfakes, and that has to do with the approach. You'll see why trying to stay ahead by developing a 'wonder weapon' is a fool's errand.

Instead, a more practical and effective solution is to outsmart fakes. You'll discover why leveraging a more comprehensive detection methodology supported by a consortium of trusted entities provides an added layer of hard-to-beat verification.

The Anatomy of a Deepfake

Anything can be broken down -- distilled into basic components. And as deepfakes go, they use two 'legs'. These legs are referred to as the injection and presentation legs -- quintessential parts for getting away with the use of almost perfect copies of originals.

Injection: How deepfakes gain access

Injection involves using deep learning algorithms like autoencoders and decoders to manipulate and generate fake content by altering existing images or videos. In a hypothetical scenario, an attacker injects a deepfake image or video into the vendor's API or software development toolkits. This action fools the vendor's systems into believing that the image or video came from a device's camera owned by a user or customer.

Presentation: How deepfakes display as convincing fakes

Presentation refers to the realistic display of manipulated content, where the generated fake content is presented in a convincing manner to deceive viewers. For example, an attacker uses their device's camera to capture the deepfake image or video. What's more perplexing is that this image or video may have been printed out or displayed on the screen of another device, showing a rudimentary but effective approach.

How deepfakes are created

Understanding the nature of fakes isn't just about knowing how they are made. It's about developing a deeper sense of key elements that are used to develop inauthentic copies. And as you'll see, detail is a major factor. The more detail available for fake-generating tools to ingest, the more convincing the fake.

Producing fakes includes several steps. Each step relies on the last, leveraging details to render a hard-to-spot fake.

Dataset collection

Large datasets of images or videos of the target are collected to form the foundation for the deepfake creation process. This data helps the algorithm learn and mimic the subject's facial expressions, mannerisms, and speech patterns. The greater the quality and diversity of data, the more realistic and believable the deepfake.

Data collection often includes identifying various poses, lighting conditions, and emotions to ensure the deepfake appears convincing. Often, various sources are used to collect data, including publicly available images and videos of the target. If need be, attackers may create their own dataset.

Preprocessing

Preprocessing involves preparing and optimizing data for the deep learning model to generate the fake. This step is essential to ensuring that the deepfake appears convincing and realistic.

Feature extraction

Feature extraction involves identifying and extracting distinctive characteristics or attributes from the source data, often related to the subject's face. Facial landmarks, such as eyes, nose, mouth, and eyebrows are collected for models to understand the geometry and structure of the face. Expressions, textures and color, gaze direction, head poses, blinking, and eye movements are also extracted.

Model training

A deep learning algorithm, often based on Generative Adversarial Networks (GANs) or other advanced models, is selected to define how the model will learn to create deepfake content. They are used to teach the AI model how to generate convincing synthetic content using data collected on the target.

Manipulation

Manipulation involves various techniques and technologies used to alter or modify source content to appear as if a target is performing it. Manipulation involves analyzing the target's facial expressions, gestures, and vocal characteristics.

Synthesis

Synthesis is characterized by generating artificial content that imitates the appearance and behavior of the target. Synthetic content is produced using deep learning algorithms like GANs consisting of two neural networks: a generator that produces fake content and a discriminator that distinguishes between real and fake content.

Realism enhancement

Several techniques and processes are used to make the generated deepfake content appear as realistic as possible. This phase focuses on refining the synthetic content through fine detailing, adjusting light and shadows to ensure the fake convincingly mimics the target.

Smoothing

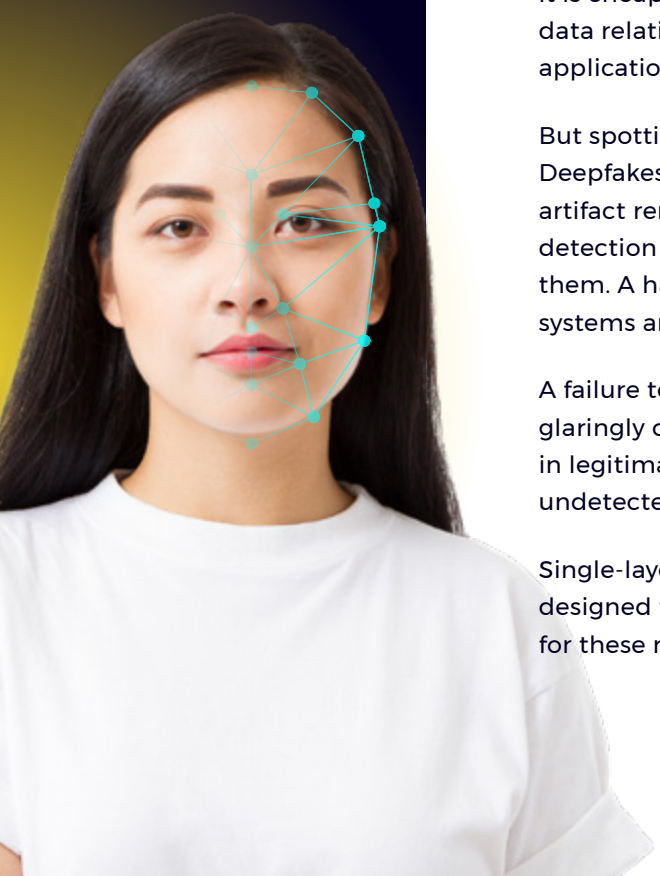
“Smoothing" refers to refining and enhancing the synthetic content to make it visually convincing and free from imperfections. Noise reduction, texture blending, and maintaining motion consistency ensure that deepfake appears natural and seamless.

Audio synthesis

Audio synthesis includes generating artificial audio that matches the lip movements and expressions of the manipulated video. This process includes voice cloning, speech generation, and emotional tone matching to produce deepfakes that sound authentic.

Visual and audio synch

Facial movements, expressions, and lip-syncing often give away deepfakes. Incorporating visual and audio syncing enables a deepfake video to match the corresponding audio, making the manipulated content appear authentic and convincing.



Post-processing

Post-processing is all about refining a polished product. A series of actions are performed on the generated deepfake to enhance its quality, realism, and believability. Actions include noise reduction for a cleaner audio-visual fake, color grading, and frame rate matching and stabilization to eliminate unnatural movements.

Rendering

As the final step, rendering combines all content, manipulated or altered, to generate the deepfake. Synthesized facial features and actions of one person are merged into the body of another. But most importantly, these separate elements must produce a deepfake that appears realistic and coherent enough to be indistinguishable from genuine content.

As complex or lengthy injection and presentation appear to be, they are not hard to perform. AI makes both processes easy to execute in as little as eight minutes.[1]

Today's Solutions

In response to the proliferation of deepfakes, the focus has been on using telltale signs that may hint at a video, image, or audio clip being inauthentic. A popular solution has been the introduction of single-layered detection.

Single-layered detection as a tool

At face value, it solved a problem. By analyzing facial transformations, discrepancies in skin texture, unnatural eye movements, irregular lighting changes from frame to frame, and other seemingly hard-to-identify elements, single AI detection caught some deepfakes.

Single-layered AI detection appears to be an attractive solution because it is cheap to deploy. And with the ability to process large volumes of data relatively quickly, it became a tool of choice for resource-limited applications.

But spotting "some" fakes has proven less than effective over time. Deepfakes have grown in complexity. Improved noise reduction and artifact removal make detecting smarter fakes nearly impossible for single-detection solutions. And that's a pressing concern for those reliant on them. A hard-to-ignore ninety-five percent of single-layer facial recognition systems are now incapable of detecting deepfakes.[2]

A failure to distinguish false positives and false negatives became another glaringly obvious shortcoming. Limited contextual understanding results in legitimate media being mistaken as a threat and real threats going undetected.

Single-layer AI detection is also easy to fool. Adversarial attacks specifically designed to outsmart single AI detectors can instantly render them useless for these reasons[3].

- **Lack of complexity:** Single-layer models have limited capacity and cannot capture intricate patterns in data. Due to their linear decision boundaries, they are vulnerable to adversarial attacks that exploit their simple decision boundaries.
- **No deep learning features:** Shallow models lack the ability to learn deep, hierarchical features from the data making them vulnerable to adversarial inputs.
- **Generalization limitations:** Single-layer models also struggle with generalizing to unseen data, a shortcoming that makes them less resilient to adversarial examples that may not conform to the training data's simplicity.
- **Limited defense mechanisms:** By design, most single-layer models often lack advanced defensive mechanisms like adversarial training or regularization techniques that can mitigate adversarial vulnerabilities.

Outsmarting deepfakes with multilayered defense

Clearly, single-AI detection systems have proven flawed. Single-AI detection's shortcomings, however, aren't in a vacuum. A larger conversation about the state of regulatory involvement in spearheading identity fraud prevention is much needed. Today, no rating systems point to how effective any identity fraud detection system is.

Accreditation, as one would observe in the hotel industry with star ratings, doesn't exist in the identity fraud prevention ecosystem. This is where the strongest, most comprehensive measures must be taken. This is why a more adept and effective solution based on a multilayered detection system is far more compelling for several reasons.

A multidimensional defense

Multilayered AI detection offers several safeguards, extending detection beyond just facial recognition elements. Combining information from multiple data modalities, such as text, audio, and video enhances the accuracy of deepfake detection by considering a broader range of features and inconsistencies. Combining more mechanisms results in a multidimensional defense that fakes must circumvent to succeed.

Beyond injection and penetration

Multilayer detection systems address the injection and presentation challenge by leveraging single-AI detection methods, but it also goes much further. It searches for behavioral clues that signify fraudulent content -- examining the fake's DNA. From facial movements and lip-syncing to micro-expressions, and more, multilayer detection scrutinizes a larger dataset to identify and outsmart fakes.

This larger dataset is also examined using error-level analysis, a way to examine inconsistencies in compression artifacts across different parts of an image or video. Error-level analysis makes spotting signs of manipulated content often found in deepfakes easier.

Advance machine learning techniques

Multilayer detection systems leverage advanced machine learning techniques, including deep neural networks and fusion methods, to continuously improve their accuracy and adapt to evolving deepfake creation methods. Adaptation is therefore accelerated, enhancing the multilayered detection systems capabilities iteratively.

A biometric layer for added verification

Multilayered AI and biometric detection deliver an even stronger and more potent layer of protection. Using biometric markers, such as facial landmarks, voice patterns, iris patterns, and behavioral biometrics of would-be targets, multilayer detection systems can spot and weed out fakes.

Scalability

Multilayered detection systems can be used across more platforms. Real-time processing with algorithms and hardware can check videos in real-time or near real-time, enabling quick detection of deepfakes in various applications. When integrated, it guards ports of entry for potential targets and creates a more secure ecosystem.

Resilience to adversarial attacks

Multilayer models are often designed to be robust against adversarial attacks, which are techniques used by deepfake creators to evade detection. These models can identify subtle artifacts and anomalies introduced during the deepfake generation process.

Human-AI blended detection

Incorporating human experts and moderators into the detection process delivers greater accuracy. While AI is exacting, human participation helps refine detection by leveraging context and intuition, elements inherently not part of AI detection systems.

Consortium-based validation

Consortium-based validation leverages data across-checking in a group of trusted entities as an added layer. This approach offers an additional fine-tooth comb to run identities through, enhancing reliability and trustworthiness through reputation scoring.

Your Best Bet at Catching Deepfakes

Several detection options are available today; however, deepfakes can only really be detected by a combination of tools. Multidimensional detection performs far more checks designed to identify fakes on a granular level. As deepfakes become more complex in nature, multilayer detection will be the most prudent approach to preventing fraud.

See How We Outsmart Deepfakes and Prevent Identity Fraud

Serial Fraud Monitor is a multidimensional detection solution based on an award-winning neural network technology that's scalable and built to create the most secure ecosystem possible for users. It outsmarts the smartest fraudsters even if they've already slipped past front-line identity verification defenses.

Consortium-based validation enhances its sophisticated and robust fraud detection features, a unique data cross-checking approach that leverages reputational scoring for stronger verification.

Features include:

1. **Fraud detection**
   Advanced neural network technology recognizes even the most sophisticated synthetic fraud.
2. **Post-breach cleanup**
   On-the-spot damage control to minimize losses and recover quickly from an attack.
3. **Traffic-level fraud analysis**
   Real-time fast reaction and insights that detect fraudulent activity based on incoming and historical patterns.
4. **Advanced machine learning and AI mechanisms**
   Highly accurate fraud detection algorithms, leveraging the latest technologies to deliver superior performance.
5. **Reputation scoring and consortium validation**
   Enhanced reliability and trustworthiness reputation scoring and data cross-checking in a consortium of trusted users.

Book your expert-led demo

Join us for a walkthrough of Serial Fraud Monitor and see how our deepfake detection technology spots deepfakes to create the safest ecosystem for your organization.

Sources:

1. AI-generated deepfakes are moving fast. Policymakers can't keep up
2. Deepfake videos easily fool face systems, researchers warn
3. A Unified Framework for Adversarial Attack and Defense in Constrained Feature Space