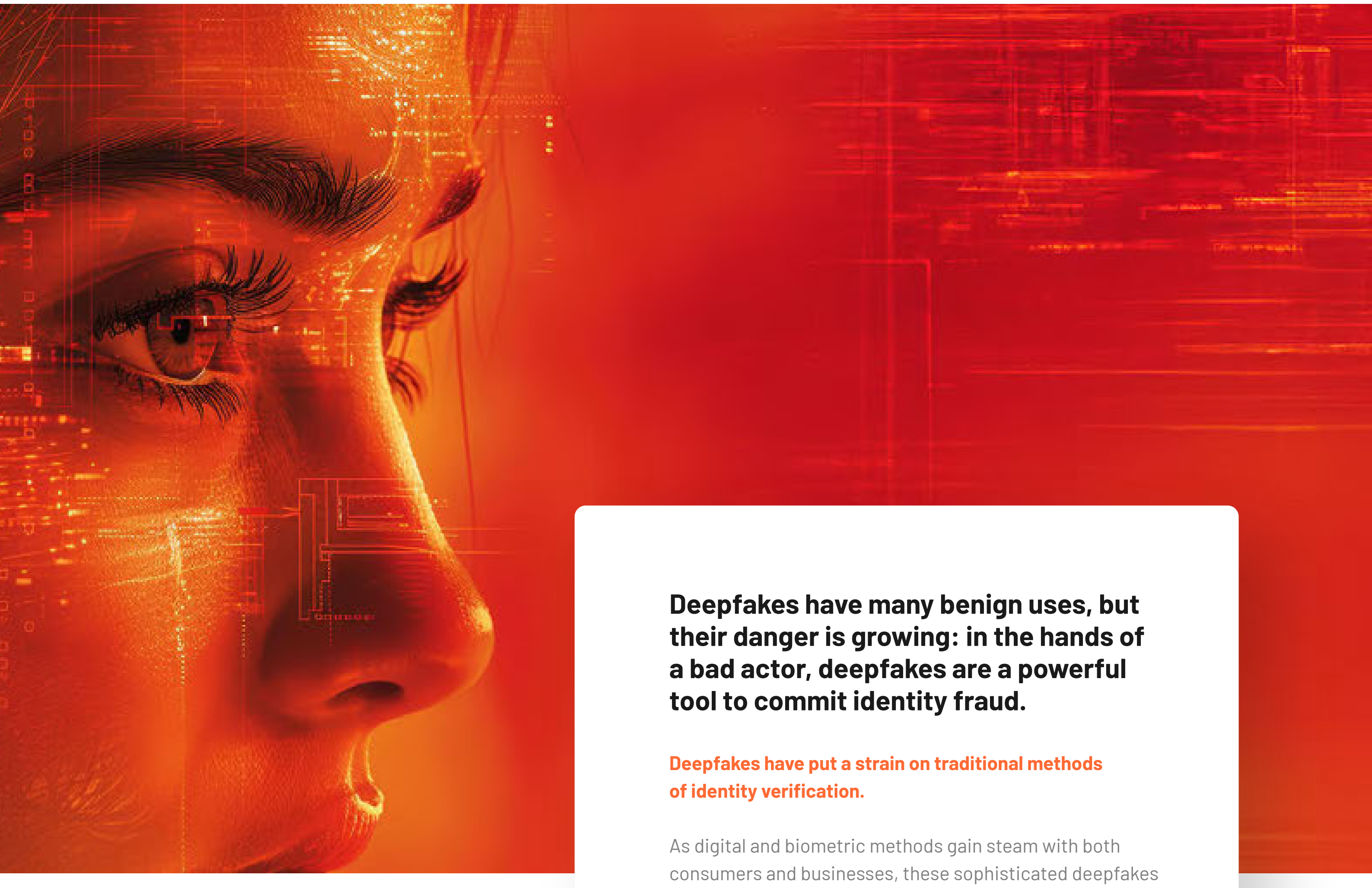
Defined & Detected: The Many Faces of Deepfakes



Deepfakes have many benign uses, but their danger is growing: in the hands of a bad actor, deepfakes are a powerful tool to commit identity fraud.

Deepfakes have put a strain on traditional methods of identity verification.

As digital and biometric methods gain steam with both consumers and businesses, these sophisticated deepfakes are often circumventing facial recognition technologies and allowing fraudsters a new tool to steal identities and commit fraud. Rates of account takeover fraud and phishing continue to skyrocket, fueled by the growing use of deepfake technology.

How can you keep your business and customers safe, and how do you identify deepfakes as they become nearly indistinguishable from reality?

ABOUT FRAUD in Partnership with AUTHENTICID

Defined & Detected:
The Many Faces of Deepfakes

Deepfakes Come in Many Forms

Fake Audio

WHAT THEY ARE:
Synthetic voice recording can mimic a person's voice and speech patterns.

HOW THEY'RE USED:
Bad actors can use fake audio to impersonate an individual for grandparent scams and more.

Fake Video

WHAT THEY ARE:
An AI-generated video which could either be an original video of a targeted individual, or a faceswap of a person onto a video of another individual.

HOW THEY'RE USED:
Deepfake videos are a prime way to spread misinformation, and they're especially dangerous when impersonating political leaders or celebrities - but they can even be used to impersonate any individual to conduct scams or bypass a biometric authentication system.

Fake Text

WHAT THEY ARE:
Any text (social media posts, articles, emails, SMS, and more) that are AI-generated to impersonate an individual and/or mislead a reader.

HOW THEY'RE USED:
Fake text can be used in phishing scams, to produce fake online reviews or create fictious stories in the news.

Fake Image

WHAT THEY ARE:
AI-generated images that can depict an actual individual or a completely fictitious subject simply from entering data into a prompt.

HOW THEY'RE USED:
Fake headshots can be used to create fake IDs, carry out identity theft, and fake images can be used to spread misinformation.

ABOUT FRAUD in Partnership with AUTHENTICID

4 Strategies for Detecting & Defending against Deepfakes

1. Assess Image/Video Features

Deepfake videos and images are getting better, but they're not perfect. When looking at videos and images, there are several telltale signs of a deepfake:

a. **Video cues:** areas around the mouth and chin may be "off." This could include fewer wrinkles, less detail, or a blurry chin.

b. **Background consistency:** While AI tools can now create convincing backgrounds, using AI tools can perform checks on image backgrounds to spot granular errors and inconsistencies that the human eye may not be able to spot.

c. **Shadows/lighting:** Deepfake algorithms still haven't gotten shadows and reflections correct. Surrounding surfaces and backgrounds, and even reflections within a subject's eyes, can be used to identify a deepfake.

d. **Extras:** Deepfake images may have extra limbs or fingers due to the prompts entered into image-generating tools focusing on names, which results in an emphasis on faces.

2. Use Liveness and Human Analysis

Deepfake images and videos are advanced enough to pass through several facial recognition technologies. But liveness detection technology can differentiate deepfakes from genuine media by detecting the "liveness" of a subject by evaluating the physiological and/or behavioral characteristics that are different from natural human movements and interactions- some that may not be detected by humans. That's why it's a crucial tool to detect these deepfakes in presentation and injection attacks.

a. **Lip syncing:** audio/video synchronization may be slightly off, which can be identified via assessing lip movements.

b. **Facial expressions and body movements:** Often, AI cannot overcome inconsistencies in facial expressions and body movements, which can be examined to identify these irregularities

c. **Pupil dilation:** If the video is high resolution, watch the pupils, which are typically not altered in deepfakes, but should change in a human subject.

d. **Strange speech elements:** Bad actors may use text-to-audio tools to recreate voices, which may not reflect normal speech patterns.

Defined & Detected:
The Many Faces of Deepfakes

4 Strategies for Detecting & Defending against Deepfakes

3. Source Analysis

The source of a multimedia file provides important clues as to whether media has been altered. Using deepfake detection algorithms can rapidly and thoroughly analyze a file's metadata for any inconsistencies or anomalies that may indicate tampering, including creation timestamps, geolocation data, and camera/device information. Even without an AI tool, confirm the credibility of any media type by analyzing the original source's reliability and reputation.

4. Methodology Analysis

a. **Presentation Attacks:** Presentation Attacks can fool biometric authentication systems by using deepfake photos or videos. These can include digital images or lifelike masks, but increasingly, they're relying on deepfake videos to deceive systems. Attackers can use this technique during customer onboarding, authentication, unlocking cell phones, and access to data or physical locations with deepfake photo or video replays and photo or video renders. In each, the key is liveness detection to determine media's authenticity.

b. **Injection Attacks:** A Digital Injection Attack occurs when a bad actor "injects" false data into an identity verification platform workflow. This is accomplished via several methods: emulators and virtual cameras can convince a system that it's receiving authentic data. In digital injection attacks, bad actors circumvent typical methods to gather identity data, like a device camera, microphone, or fingerprint sensor. An attacker may also use false location data. Deepfakes are an important tool in these attacks, as convincing headshots, selfies, and IDs can be inserted into these processes with relative ease, and they're increasingly difficult to detect. Via algorithms and liveness detection, such as AuthenticID's recently announced solution, these attacks can be minimized.

c. **Deepfake Detection Techniques:** Face Swaps, GANs, and Reenactments
– GANs: Deepfakes often use Generative Adversarial Networks (GANs), involving two neural networks: a generator and a discriminator. The generator creates images by analyzing source content and extracting features, while the discriminator evaluates and corrects these images. Repeated iterations improve image fidelity. GAN-based deepfakes can be lower quality due to blending artifacts but can also be trained to detect deepfakes.
– Face Swaps: This method replaces a face in a video with an AI-generated one using generative autoencoders. It encodes and decodes images through multiple layers, maintaining attributes like expression and orientation. Autoencoders provide accurate swaps as they don't insert missing data.
– Reenactments: Also known as puppet-master deepfakes, this technique transfers facial expressions from a source performer to a target face. It manipulates facial expressions in images or videos, often leaving detectable signs of forgery.