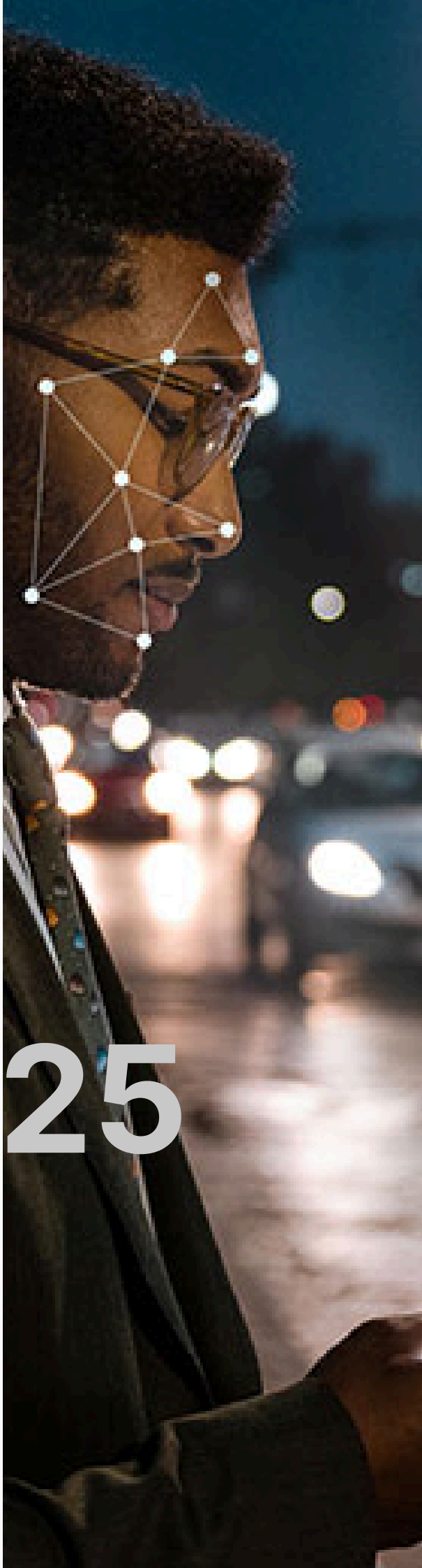
authID

DEEPFAKES

COUNTER MEASURES

2025



Help eliminate fraud and account takeover with market-leading biometric identity authentication

Table Of Contents

White Paper

DEEPFAKES
COUNTER-MEASURES

Just as the internet itself transformed our world, Artificial Intelligence (AI) is guiding our digital universe to completely new heights. AI-based Large Language Models (LLMs) can ingest huge amounts of information, calculate the relationships between words and data points, then write text, answers to complex questions, solutions to math problems, even generate computer code. AI can predict cancer, plot routes for vehicles, spacecraft, and commerce. AI can detect fraud by looking at people, their profiles, their activity, their backgrounds, their identities. AI makes itself exponentially smarter, exponentially faster than human programmers ever could. It has taken on a life of its own, through adversarial networks in which opposing digital entities help each other train by bouncing opposing datasets off each other using Generative Adversarial Networks (GANs).

While GANs are training Large Language Models (LLMs) to generate better phishing emails and in greater volume, generative AI is also improving in creating visual and audio imposters in the form of one of the most frightening new forms of fraud: deepfakes.

DEEPFAKES
THE NEXT GENERATION THREAT

Deepfakes can be synthesized from a starter photo, such as one acquired from LinkedIn, Facebook, or any other common source, then repurposed. A still photo can be made to blink, smile, nod, or take other actions to mimic a live person. They can even be made to speak. Entire videos can be generated from still photos.

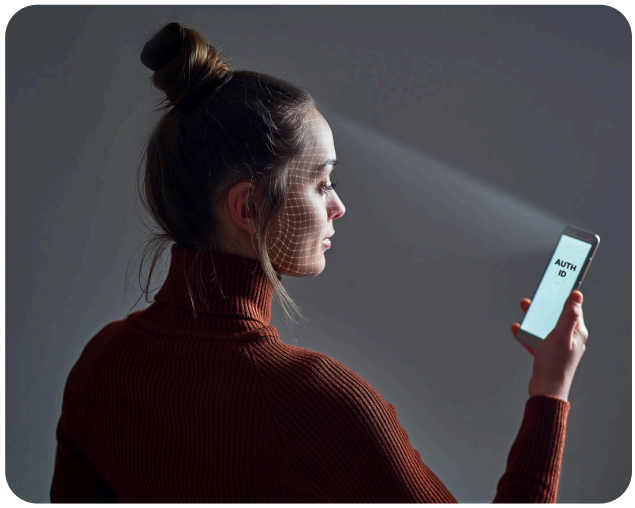
While they can be used for entertainment, education, and other legitimate purposes, deepfakes are increasingly leveraged for illegitimate purposes. Deepfakes have been utilized in disseminating misinformation, propagate vulgar material, and meddle in politics. They are also increasingly used for committing fraud. A common example of this is how deepfakes are purposed by criminals to fool validation systems. Phony faces, realistic-looking and even animated as if live, are presented within an onboarding flow, pretending to be a stolen identity, or even a totally synthetic identity. To augment the phony faces, deepfake technology can also generate documents, such as physical ids. These might be drivers licenses, passports, or any other kind of identification.

According to a study by Signicat, while deepfake attacks constituted less than 1% of fraud attempts in 2021, that number rose to 6.5% by 2024, an increase of over 2000%.

Source: Signicat (“The Battle Against AI-Driven Identity Fraud”)

In an example of a highly sophisticated deepfake attack, an officer of a Hong Kong-based organization was tricked into wiring $25 million to multiple account over multiple transactions during a video conference in which the participants appeared to look and sound like executives of his company but who were in fact all deepfakes.

Such a crime required a high degree of orchestration and preparation, while daily occurrences of fraud are empowered by deepfakes, and with far less talent required.



In the past, traditional hackers possessed deep knowledge of networks and other digital systems. They were replaced by “script kiddies” who needed at least some ability to wield downloadable intrusion kits. But the advent of AI tools have enabled an entirely new generation of criminals who only need to prompt machine learning models to create deepfakes on their behalf.

A common procedure to apply for online accounts is to provide a physical id document along with a selfie. Therefore the combination of deepfake ids and faces supplies fraudsters with everything they need to acquire profiles, credit cards, and other instruments of ill-gotten gains. And again, it can be a simple matter to create deepfakes. But the method of their delivery may not be nearly as simple.

HOW ARE DEEPFAKES USED IN STAGING AN ATTACK?

There are two basic attack vectors for criminals who wield deepfakes. One is simple, requiring a minimum of talent, while the other requires additional abilities, although it too is becoming easier to pull off. These two vectors are Presentation and Injection.

PRESENTATION ATTACK

Simply enough, this entails putting an improper or fake image in front of the camera. This could be a modified image, a deepfake, a printout or other picture, or a screen replay, rather than a live, real-time id document or person. This has been a method for as long as there has been digital onboarding. But the level of sophistication has definitely increased. A variety of tools are available to help criminals create fake documents with images that match the person presenting them (if in fact that person is real). Fraudsters still regularly use common apps such as Photoshop and even MSPaint to manipulate images.

But these kinds of manipulations are one-offs, with each one taking plenty of time. Deepfakes can be manufactured quickly and in high volume. AI can generate a fake id with necessary information, image, even bar code, along with a facial image that matches the portrait on the id, so that they two can be presented in tandem.

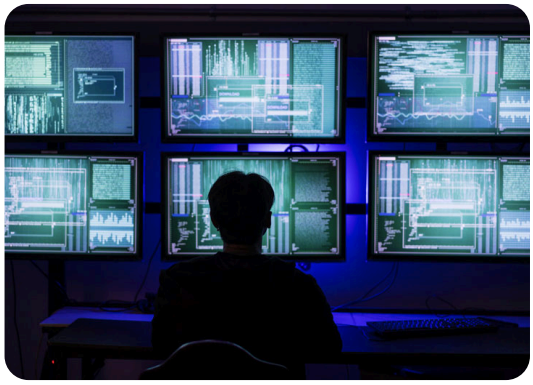
While talented criminals can manipulate images with more traditional tools to the degree that many of them pass detection, deepfake tech lets those same criminals more quickly and easily perfect modified or fabricated images that are even more likely to succeed in fooling both digital and human defenses. Organizations that employ manual review, meaning the use of humans to examine submitted images, are subject to statistics indicating that less than one percent of people can accurately distinguish fake content from real.

The fakes still need to be presented physically, which remains a bottleneck. There are also a variety of technologies which organizations employ for detecting fakes at the point the images are presented to the camera. These two limitations can compel fraudsters to employ the second kind of attack.

INJECTION ATTACK

Whereas presentation attacks by themselves have a single attack surface, namely the camera, injection attacks enjoy a variety of routes they can take in delivering deepfakes to the target. Injection attacks work by going around the camera. Images are inserted into the pipe to the backend server behind the expected sensor.

- Virtual cameras
- Additional hardware
- Man-in-the-middle attack
- API corruption. Images are typically transported by API calls after being transformed into Base 64 format. If the fraudster can determine how to spoof the expected API call, they can create their own Base 64 payload and post it to the back end.



Injection presents two advantages for the attacker to get past the weaknesses of simple presentation:

1. It allows attempts in even greater volume, since it is not limited by the processing time required by image capture and can be launched via automation.

2. It bypasses the upfront defenses that would otherwise detect the fakes at the point of capture.

Just as with missile defense, an injection only has to be right once, while the defender must be right all the time. With so many options for launching an injection attack, a deepfake-powered fraudster can take multiple paths to their goal.

Presentation attacks represent up to 12% of fraud attempts, while injection attacks represent 7.5%. In the financial services sector, AI-driven assaults are swiftly approaching half of all fraud attempts.

HOW TO DEFEND AGAINST DEEPFAKES

So with the newly-minted might of Artificial Intelligence behind these deepfake attacks, what are the options available to those who want, and absolutely need, to fight back?

The different attack vectors require different defenses that meet them where they stage their assaults. Presentation attacks appear at the camera, so that's where to start. Injection attacks seemingly have myriad paths, so they seem more problematic, therefore the counter-measures need to be simplified. We can't be everywhere, but we can be standing, rolling pin in hand, where they intend to go. Let's examine these two assaults on their own merits.

HOW TO PREVENT PRESENTATION ATTACKS (AND HOW TO FAIL)

One could say that any fake, whether created with Photoshop or through AI, is a candidate for detection. It's not real, correct? The issue with deepfakes is their twin threat of presenting superior quality fakes, and in volume (given the ability to manufacture them instantaneously).

First, manual review. This traditional method involves routing images of ids and people through a boiler room of examiners who will allegedly pick out the fakes from the real ones. Granted, one thing that humans can do better than many digital systems is spot fraudsters wearing masks that are meant to either resemble the person on the id, or simply mimic the mathematics of the face (distance between eyes, nose, chin, lips, etc.) that many digital detectors rely on. But beyond that, human detection is terrible. It may be obvious why, but let's state the obvious:

- Statistically humans do a lousy job of spotting deepfakes. As previously noted, they succeed at such detection less than one percent of the time.
- They don't scale. Human reviewers are a bottleneck. Unlike digital systems, they don't react at the speed of thought. Many crypto sites still utilize this method, making them vulnerable to fraud, while keeping applicants waiting for literally weeks to be approved.
- Routing personal documents past human reviewers means that your documents are in the hands of strangers, along with names, addresses, dates of birth, and actual document numbers. Where will these end up?
- Therefore manual review is appropriate for very small operations.

The second approach to spotting deepfakes is liveness. This is the ability for a digital onboarding system to determine that an image put in front of the camera is being done so by a live human, and in real time. If you are taking a picture of an id document such as a license or passport, is it present at that moment, or is it something else? Is it a picture, a printout, a screen replay, a video? Liveness detection is designed to determine this immediate presence, with a carbon-based entity showing their document and/or their face at that time.

NIST publishes evolving standards for liveness detection, stating that it should be accurate to within 1-to-100K attempts. But even meeting this standard is a numbers game. With the sheer volume of AI-generated, digitally-manufactured attacks possible, that rubric means a lot of fakes will still get through. Too many detection platforms hit that "performance" level and call it a day.

There is more than one type of liveness detection. A popular one is active liveness. This involves making the user blink, nod, smile, even rotate their head while looking at the camera. In other words, tell them to look alive. On the face of it, this seems to be a great way to detect that the individual taking their picture is not themselves a picture. But active liveness has inherent weaknesses:

- **Using free software, fraudsters can animate a still photo, making it nod, blink, smile, rotate, and even speak**
- **Users typically dislike having to perform ridiculous movements to verify themselves.**
- **Users sometimes dislike the fact that their own face appears in the camera view, as it feels like an invasion of privacy, despite knowing they are submitting a picture.**
- **Verification takes longer, as there is motion data to process.**

The other option is passive liveness. The user simply takes their own picture, without moving their head around and performing various gestures. Processing time is decreased. More sophisticated systems can still determine liveness by using one or more still frames. Smartphone users are accustomed to authenticating with face id systems which allow them instant access to their phones, although those capabilities are also subject to being fooled with pictures or deepfakes, and only get them as far as their device desktop or phone apps. This isn't liveness so much as basic facial recognition.

In the case of liveness, it's not so much what you do as how well you do it. Basic liveness detection can still fail against the most polished of deepfakes.

An additional consideration is this: even if a physical id is deemed to have been presented live, that does not mean it is legitimate. Deepfaked (or traditionally-faked) ids can be printed, laminated, and presented live to the camera. So further examination is still warranted.

HOW TO PREVENT INJECTION ATTACKS (AND HOW TO FAIL)

If liveness detection and other image verification methods are the gate, criminals might simply choose to go around the gate and insert images on the other side of this first hurdle, utilizing any one of the aforementioned routes through hardware, software, or networking hacks. In a totally homegrown environment, where the organization controls every possible route, it might be possible to track all the traffic through all the possible doors, but missing even one door leaves an open path for intrusion. And in an era where companies and agencies assemble their platforms with best-of-breed options, using over-the-counter software and hardware, that total control is not possible. This makes the task of watching all the paths and doors very difficult.

Many verification systems identify cameras by naming them. But this is ridiculously simple to defeat. Payloads can be encrypted, but encryption does not validate the payload's origin. Code tampering, such as when injecting an image directly into an API call, can be defeated by code obfuscation, but this does not prevent the fraudster from the images from being manipulated before the code receives them.

So what is the answer to covering all those digital highways? Watch the destination. Did the image that arrived at the server, ready to be validated, originate from where it was supposed to? If it shows up without the right credentials, so to speak, it's not valid. This means coordination of a kind between the front end and the back. The server side needs to know what the front end is sending, with a type of signature. In this way, the final payload comes with a star of approval, indicating its legitimate provenance.

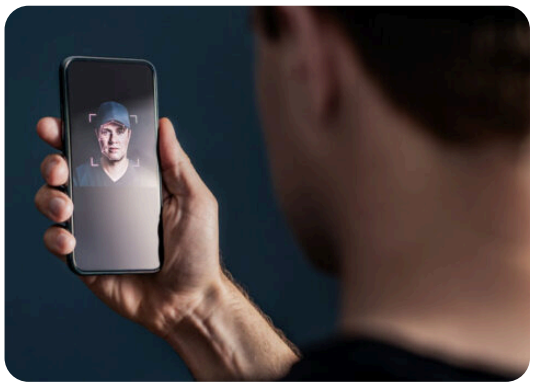
The integrity of the originating device is also in play. Has the phone or desktop been compromised? Is there additional hardware plugged in that imperils the legitimacy of the image capture? Have software tools been employed to inject images into the camera itself without an actual picture being taken?

All of this seemingly simplifies the process of validating the capture, but there are clearly multiple ways to corrupt the device of origin, and few providers account for all these possibilities.

AuthID's MULTI-LAYERED APPROACH TO DEEPFAKES

As security professionals with deep expertise in fraud, authID's technical leadership goes further in uncovering deepfakes than any other organization. We recognize the promise as well as the threat of Artificial Intelligence, including the extreme capabilities of deepfakes for committing fraud. Our ProofTM solution for onboarding consumers and workforce combats deepfakes to ensure that only legitimate users are able to create accounts, while our VerifiedTM solution further ensures that only registered users can leverage their biometric root of trust to authenticate on a daily basis.

ProofTM captures images of physical id documents as well as **selfies**, validates them discretely, then matches the images to triangulate a legitimate individual before creating that root of trust. **VerifiedTM** accepts facial images, matches them with either a biometric hash or private key to verify them, then allows them access. Both solutions utilize our cutting-edge technology for detecting deepfakes. In this way, phony ids and faces cannot register, nor can they leverage existing accounts to steal access.



How Do We Definitively Defeat Presentation Attacks?

We start with passive liveness detection, to determine that the id as well as the person in front of the camera are in fact present, in real time. We detect printouts, screen replays, and videos. Most importantly, our market-leading technology examines both the visible and invisible artifacts present in deepfakes.

How Do We Definitively Defeat Injection Attacks?

First, we examine the integrity of the device. We ensure it has not been tampered with, that no malware or other agent is present that would compromise the device performing the image capture. We check for virtual cameras and other hardware intrusion.

The resolution between physical and virtual cameras can be different, as are the API functions. Third-party browser plug-ins or malicious JavaScript can also accomplish injections.

Our algorithms are trained on detecting even the most sophisticated deepfake images, finding inconsistencies and anomalies in movements, image textures, and other indicators of manipulation or fabrication.

This multi-layered methodology to deepfakes cuts off all standard avenues for cyber-criminals seeking to corrupt the onboarding or authentication process with AI-generated images or videos. Unlike many vendors, authID has been ahead of the deepfake curve for years, and does not fall back on traditional approaches in which older technologies are optimistically applied to a newer and ever-evolving threat. Instead, authID has invested strategically in constantly enhancing out platform, to keep our clients safe from the most malicious uses of Artificial Intelligence with our own forward-looking solutions.

About AuthID

authID (Nasdaq: AUID) ensures enterprises “Know Who’s Behind the Device™” for every customer or employee login and transaction through its easy-to-integrate, patented, biometric identity platform. authID quickly and accurately verifies a user’s identity and eliminates any assumption of ‘who’ is behind a device to prevent cybercriminals from compromising account openings or taking over accounts. Combining secure digital onboarding, biometric authentication, and account recovery with a fast, accurate, user-friendly experience, authID delivers biometric identity processing in 700ms. With our ground-breaking PrivacyKey Solution authID delivers all the benefits of biometric identity verification, with a 1-to-1-billion false match rate, while storing no biometric data. Binding a biometric root of trust for each user to their account, authID stops fraud at onboarding, detects and stops deepfakes, prevents account takeover, eliminates password risks and costs, and provides the fastest, most frictionless, and most accurate user identity experience demanded by today’s digital ecosystem. Contact us to learn how authID can transform your organization’s fraud defenses, streamline user authentication and ensure you always “Know Who’s Behind The Device.”