Emerging Tech: The Impact of AI and Deepfakes on Identity Verification

Published 8 February 2024 - ID G00803486 - 17 min read

By Analyst(s): Swati Rakheja, Akif Khan

Initiatives: Emerging Technologies and Trends Impact on Products and Services

Deepfakes generated using generative AI technologies pose a fundamental threat to the integrity of identity verification. Identity verification product leaders must understand this emerging threat and take a proactive approach to differentiate and secure their solution offerings.

Overview

Key Findings

- Liveness detection technologies are becoming critical for defending against deepfakes and verifying the genuine presence of an individual user during the “selfie capture step” of the identity verification process. This is driving vendors to use a combination of multiple factors to differentiate their solution offerings and offer more comprehensive protection.
- Recent advancements in generative AI (GenAI) are making deepfakes increasingly sophisticated and adaptable as advanced attackers can now mimic facial expressions, blinking patterns and even subtle micromovements with uncanny accuracy, confounding even the most advanced detection algorithms. Product leaders in the identity verification space are being driven to adopt a more holistic approach that incorporates a multilayered defense strategy to defend against deepfakes.
- Deepfake attackers are weaponizing the rapid evolution of GenAI, constantly inventing new and more sophisticated attack techniques. As GenAI continues to rapidly evolve, identity verification product leaders will need to actively engage with AI and security experts to anticipate future attack vectors and proactively develop countermeasures.

Recommendations

To defend against these rising deepfake attacks, identity verification product leaders must:

- Invest in the development and implementation of a combination of active and passive liveness detection strategies to assess genuine presence and detect deepfakes, with a strategic focus on passive liveness detection as AI tech matures and attacks become more sophisticated.
- Integrate detection capabilities for additional signals indicating an attack, choosing between in-house development or partnerships/mergers and acquisitions (M&As) with existing vendors by evaluating the level of product maturity and commoditization.
- Invest in a threat intelligence team focused on tracking emerging deepfake-related threats and collecting intelligence on various techniques being used by attackers. Additionally, product leaders should leverage GenAI to their benefit, using synthetic data to strengthen machine learning (ML) algorithm training.

Strategic Planning Assumption

By 2026, attacks using AI-generated deepfakes on face biometrics will mean that 30% of enterprises will no longer consider such identity verification and authentication solutions to be reliable in isolation.

Analysis

Technology Description

GenAI technologies can generate new derived versions of content, strategies, designs and methods by learning from large repositories of original source content. GenAI can have profound impacts on various aspects of business, including content discovery, creation, authenticity and regulations; automation of human work; and the customer and employee experience (see Emerging Tech: Primary Impact of Generative AI on Business Use Cases).

The advent of GenAI has increased the sophistication of attacks that identity verification vendors must defend against. GenAI tools are capable of producing seemingly real content in voice, video and image format with minimal technical input, and deepfake misuse can subvert the verification process. Though deepfakes have existed for some time, the proliferation of user-friendly tools has made their creation more readily accessible, even to individuals with limited technical proficiency. The number of deepfakes detected worldwide in 2023 [1] was 10 times the number detected in 2022. Gartner estimates the time to reach the early majority (i.e., more than 16% target market adoption) for deepfakes is one to three years because deepfakes go hand-in-hand with the GenAI advances that underpin their creation. (See Emerging Tech Impact Radar: Artificial Intelligence). This requires that identity verification vendors take a multipronged approach to safeguard against these rising deepfake attacks.

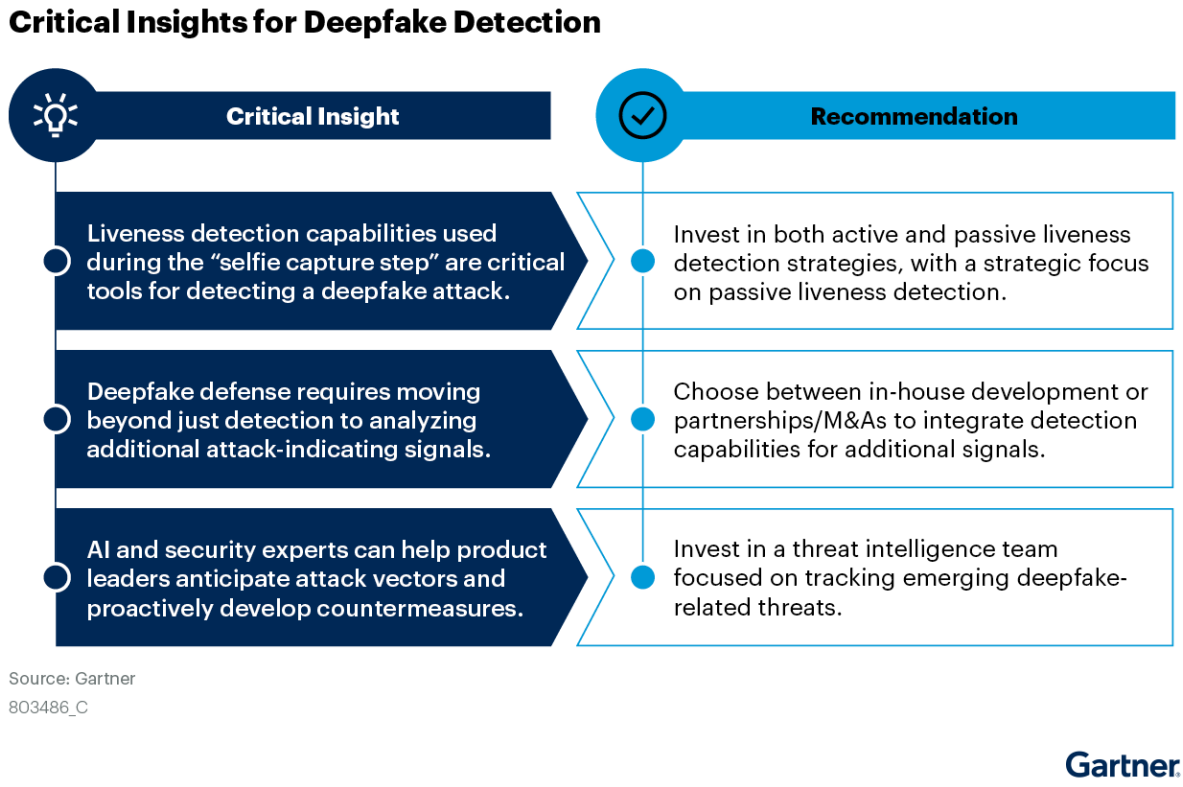
Market Definition

Gartner defines identity verification as the combination of activities during a remote interaction that brings a real-world identity claim within organizational risk tolerances. Identity verification capabilities, delivered as SaaS or on-premises, provide the assurance that a real-world identity exists and that the individual claiming the identity is its true owner and is genuinely present during a remote interaction.

This typically involves a person capturing a real-time image of their photo identity document, which the tool inspects for signs of counterfeit or forgery. Once the authenticity of the document is established, the person is prompted to capture a photo or a video clip of their face. During this step, the tool establishes the genuine presence of the person using liveness detection (or, formally, presentation attack detection), followed by biometric facial comparison with the photo from the identity document.

Some identity verification vendors also capture voiceprint during the verification process, to be leveraged in the future for the purposes of contact-center voice authentication. That user flow is susceptible to voice deepfakes, and some vendors are investing in real-time voice deepfake detection capabilities as well. However, for the purpose of this research note, we have focused on deepfakes targeting selfies or the video capture process during identity verification. Figure 1 summarizes the critical insights for deepfake detection discussed in this document.

Figure 1: Critical Insights for Deepfake Detection



Critical Insight: Liveness detection mechanisms have become critical to subvert deepfake attacks.

During the identity verification process, attackers may use deepfakes to target either the facial matching step or the document verification step, or both, in two ways:

- **Presentation attacks** — In which the attacker uses their device's camera to capture the deepfake image or video that may have been printed out or is being displayed on the screen of another device.
- **Injection attacks** — In which the attacker directly injects the deepfake image or video into the vendor's API or software development toolkits (SDKs), fooling the vendor's systems into believing that the image or video came from the device's camera.

Presentation attacks are easier for attackers to carry out. Gartner's discussions with both end-user organizations and vendors have indicated a sharp increase in the number of such attacks in the last year as GenAI technologies and tools have captured public attention. Such attacks are also easier to detect since many vendors can determine if an image is being taken of another device's screen, although this becomes harder when ultra-high-resolution screens are used. Injection attacks, on the other hand, are harder to carry out since they need more technical expertise, but they can also be harder to detect.

In terms of technical capabilities, liveness detection can be an important tool to detect a deepfake. This technology is used by vendors to assess the genuine presence of the individual during the "selfie capture step" of the identity verification process. Active techniques rely on the user having to take some action during the selfie capture process, such as turning their head as instructed or smiling. This assessment can be further strengthened by introducing randomized, surprise actions that the user must take to confuse the attacker. This can help avoid scenarios in which the attacker, familiar with the selfie capture process and associated liveness detection requirements, presents a prerecorded deepfake video. The attacker would instead need to create a deepfake in real time to respond to the randomized challenge. Passive techniques on the other hand may involve looking for micromovements in the face, 3D depth analysis and changes in light reflection as blood flows under the skin.

There is no clear evidence regarding whether active or passive liveness detection is better-suited to detecting deepfakes. There are examples of both active and passive liveness detection solutions meeting Level 2 certification by iBeta against the International Organization for Standardization (ISO) 30107-3 standard, but it should be noted that the scope of the standard does not cover injection attacks. The same is true of the recent National Institute of Standards and Technology (NIST) Face Analysis Technology Evaluation focusing on PAD. On the one hand, active liveness detection forces an attacker to create a video rather than a single image, and it introduces temporal and spatial artifacts that aid deepfake detection. On the other hand, with passive liveness detection, attackers cannot easily understand how liveness is being assessed and are therefore unable to generate a deepfake attack specifically tailored to defeat the detection process.

Near-Term Implications for Product Leaders

Vendors offering identity verification solutions should expect challenges and concern from both existing clients and sales prospects regarding the viability of their solutions if presented with counterfeit/tampered documents or deepfake images, video and/or audio. It is likely that clients and prospects would want to discuss what types and frequency of such attacks you are experiencing today. This presents an opportunity for differentiation in the crowded market through thought leadership (e.g., blogs, whitepapers, webinars) acknowledging the issue and explaining the current state of the art in terms of mitigation. Vendors that fail to explain the accuracy of their solutions in more than just simplistic terms may find themselves at a disadvantage to their competitors who do.

From a technical capability perspective, both active and passive liveness detection techniques have pros and cons pertaining to their impact on user experience (UX) and their accuracy in establishing a genuine user presence. Given the pace at which deepfake technology is improving, vendors should assume that active liveness detection will become more susceptible to attack since it could be possible to replicate the requested action gestures in real time as GenAI tech advances. Thus, passive detection may be a more strategic approach in the future. However, deploying both active and passive techniques may be more prudent in the short term.

When it comes to the specific use case of synthetic document images, the use of NFC for document verification can be a more accurate way to assess document authenticity. But this comes with its own set of challenges related to the limited adoption of chip-enabled documents and of NFC-enabled smartphones, as well as the impact on UX from the need to download a mobile app.

Recommended Actions

- Invest in liveness detection to assess genuine presence when presented with a deepfake, with a strategic focus on passive liveness detection.
- Do not rely solely on iBeta or NIST testing to demonstrate your efficacy with respect to liveness detection. Consider these to be baseline qualifications only. Demonstrate to your clients that you test your liveness detection using a broader and more robust set of attack vectors involving deepfakes and injection attacks.
- Decide whether your strategy will consist of developing deepfake detection capabilities in-house, with the ongoing investment that will be required given the rapid pace of GenAI advancement, or whether you should look for possible vendors who focus on liveness and deepfake detection.

- Proactively acknowledge the challenge from deepfakes and educate your customers about your defense strategy against these attacks.

Critical Insight: Broader defense against deepfakes requires the use of multiple signals indicating an attack.

Today many fake, AI-generated images on deeper inspection lack clarity around finer features such as hands, eyes and teeth, though they are rapidly improving. Flaws can be detected using computer-vision ML models that can spot subtle but anomalous identity features across different faces, such as strands of hair in identical configurations across multiple fake photo or document submissions. Other ways to detect a deepfake include looking for a lack of natural movements such as blinking or natural elements such as shadows. In authentication use cases, models can even compare a speaker's facial movement against prior video instances.

However, rapid advancement in AI raises a strong possibility that the technology will reach a point where it may not be possible to detect deepfakes through these techniques, or with the combination of passive and active liveness detection described above. In this scenario, a more reliable approach for identity verification product leaders would be to expand their defense beyond deepfake detection to look at a combination of signals that can indicate an attack, correlating and scoring across layers of context and device data. This strategy helps ensure that even if the deepfake itself goes undetected by the system due to its high authenticity, the attack will most likely be detected.

Table 1 lists a number of additional capabilities that can be used to detect an attack.

Table 1: Capabilities Used to Detect Deepfake Attacks

(Enlarged table in Appendix)

Capability	Description	Application
Device profiling	Running metadata gathered from a device's hardware (CPU, GPU, screen resolution) and software (OS, browser, language, time zone) through deterministic rules to identify anomalies	Both presentation and injection attacks
Behavioral analytics	Session-tracking capabilities that monitor user interactions with the protected service to build trust models for distinguishing genuine users from bots	Both presentation and injection attacks
Location intelligence	Leveraging a multitude of signals gathered from the device and environment to fingerprint a particular location beyond what is reported by GPS or IP addresses to detect fraudulent behavior	Both presentation and injection attacks
3D image detection	Detecting the absence of a 3D presence indicating that the 2D image being presented could be a deepfake, printed or projected from a screen	Presentation attack
Screen detection	Detecting the presence of glare or reflections that can be signs that a screen is being used to display an image or video to the camera	Presentation attack
Emulator detection	Detecting virtual cameras and also virtual mobile devices being run on larger machines	Injection attack
Metadata inspection	Detecting deviation from the vendor's expectation for a device's camera in terms of, for example, image size and resolution. An image or video that does not conform to the expectation may have been injected.	Injection attack
Telltale signatures	Adding watermarks to the images or videos being taken via their API or SDK. Any image or video that does not contain these telltale signatures may have been injected into the workflow.	Injection attack
API/SDK payload integrity	Cryptographic signing of the payload (containing the image or video) being sent from the vendor's API/SDK to its servers, which helps to prevent injection attacks at the network level as opposed to the device level	Injection attack
Data affirmation	Extracting data from the identity document and then checking it against another source such as an identity graph, a credit bureau or a government issuing authority	Both presentation and injection attacks
Human analyst	Using human analysts for image inspection in case ML/AI model does not generate a high confidence score	Both presentation and injection attacks

Source: Gartner (January 2024)

Near-Term Implications for Product Leaders

On a broader level, identity verification vendors must not hinge their defense against deepfake attacks purely on active/passive deepfake detection but should also monitor additional signals that can indicate an attack such as device profiling, behavioral analytics and location intelligence. However, it may not be prudent to aim to develop all these capabilities in-house. For some capabilities, such as device intelligence, device profiling and location intelligence, partnering with an established industry vendor may be wiser in terms of resource and budget prioritization. Other capabilities, such as screen detection, emulator detection, telltale signatures, layered scoring context and API/SDK payload integrity, are not as easily commoditized and will need in-house development effort.

Of all the above-mentioned signals, human analysts can add value on the “image inspection” layer while all the other layers (liveness detection, device/location, metadata, telltale signatures, etc.) generate signals that an ML model would need to interpret. Theoretically, humans can play a significant role in detecting deepfakes. However, the hybrid option offered by many identity verification vendors, which uses a human analyst if the ML algorithm cannot generate a high enough confidence score, may not be as effective as one might expect.

Thus far, most of the multiple research studies conducted on this subject suggest that humans can spot a deepfake with the same level of accuracy as an ML model, if not greater accuracy. Humans and AI tools have their respective strengths. [2] For example, AI is better at detecting nuances or patterns in deepfakes, while humans tend to rely more on the contextual signals such as why the subject in the video would behave in a particular fashion. [3] Humans may also be better at recognizing or questioning the moral or political motives behind a fake video.

While these can be powerful capabilities for discerning fake content in the public realm, such contextual signals are absent in most use cases targeted by identity verification vendors. Thus, as technology advances and deepfakes become more sophisticated, the role of human analysts in deepfake detection could be diminished.

Recommended Actions

- Don’t rely on just being able to detect fraudulent documents or deepfakes themselves. Invest in adding signals such as device profiling, behavioral analytics and location intelligence — these signals could be excellent context for detecting attacks, even if the fraudulent document or deepfake itself remains undetected.
- For the different capabilities listed in Table 1, choose between in-house development — with the ongoing investment that will require for constant retraining of AI modules — or assessing the market for possible partners/acquisition targets that are developing expertise in the capability. To some extent, the choice will depend on the capability’s level of commoditization.

Critical Insight: Understanding how deepfakes are being created by attackers is critical to stop emerging threats.

GenAI is a rapidly advancing technology, and the broad availability of new GenAI tools requires that vendors make proactive efforts to remain one step ahead of attackers. One important step in this regard is investing in a threat intelligence unit or team focused solely on gaining an understanding of the latest tools and techniques being leveraged by attackers to generate deepfakes and bypass established checks. Understanding the actual tools and tactics, techniques and procedures (TTPs) being leveraged by attackers can allow product leaders to stay abreast of the latest attack patterns and strengthen their own defense by training their detection algorithms in an accelerated manner. Vendors can understand the signatures from certain model techniques and build this into their detection capabilities.

GenAI's ability to develop synthetic datasets can give product leaders an interesting way to use the technology in their defense. By reverse-engineering the attack variants, synthetic datasets mimicking the emerging attack patterns identified from above-mentioned efforts can be used to tune the algorithms for better detection rates.

Outside a strict security context, GenAI can also be used to help identity verification product leaders address the issue of demographic bias in face biometrics processes. A challenge that many vendors face is obtaining large datasets of faces on which to train the ML models in their biometric platforms. Keeping production images of users for ML training presents a legal and business minefield. Buying large datasets is an option that many vendors pursue. The challenge with either approach is how to generate a training dataset that is genuinely diverse with respect to demographics such as gender, race and age. A lack of diversity in training data can result in ML algorithms that exhibit bias. The creation of deepfake images using GenAI is one solution to this challenge. Large datasets of synthetic faces can be created with artificially elevated levels of training data for less populous demographic groups to create a better representation of diversity. This can lower the cost and effort involved in obtaining datasets, and also help to minimize the introduction of bias into the biometric processes.

Near-Term Implications for Product Leaders

Attackers are leveraging GenAI to launch sophisticated attacks at unprecedented scale, and it has become increasingly difficult for vendors to thwart these attacks solely with manual defensive approaches. Proactive threat-hunting can help vendors stay ahead of the attackers. This can include investing in bug bounty programs to reward anyone who alerts the organization about ways to bypass the guardrails with fake content.

Another important response strategy can be to establish dedicated teams focused on engaging with your customers. Such efforts to foster direct and dynamic lines of communication can help product leaders get a continuous view of the evolving nature of the threats that customers are encountering on the ground.

Recommended Actions

- Invest in teams focused solely on tracking the latest available tools and techniques for creating deepfakes, in order to maintain the most up-to-date perspective on potential attacker capabilities and your ability to detect them.
- Use GenAI and deepfakes for positive purposes by creating datasets of faces for training your ML models. These can be designed to maintain ideal distributions among demographic groups, helping to minimize bias in your biometric processes.

Evidence

[1] Sumsub Research: Global Deepfake Incidents Surge Tenfold from 2022 to 2023, Subsum.

[2] Deepfake Detection: Humans Vs. Machines, arxiv.org.

[3] Deepfake Detection by Human Crowds, Machines, and Machine-Informed Crowds, PNAS.

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

Market Guide for Identity Verification

Emerging Tech Impact Radar: Security

Emerging Tech: Security — How to Stay Relevant as an Identity Verification Vendor

Emerging Tech: Top 4 Security Risks of GenAI

Generative AI: The Basics (Shareable Slides)

Emerging Tech Impact Radar: Online Fraud Detection and Prevention

Table 1: Capabilities Used to Detect Deepfake Attacks

Capability	Description	Application
Device profiling	Running metadata gathered from a device's hardware (CPU, GPU, screen resolution) and software (OS, browser, language, time zone) through deterministic rules to identify anomalies	Both presentation and injection attacks
Behavioral analytics	Session-tracking capabilities that monitor user interactions with the protected service to build trust models for distinguishing genuine users from bots	Both presentation and injection attacks
Location intelligence	Leveraging a multitude of signals gathered from the device and environment to fingerprint a particular location beyond what is reported by GPS or IP addresses to detect fraudulent behavior	Both presentation and injection attacks
3D image detection	Detecting the absence of a 3D presence indicating that the 2D image being presented could be a deepfake, printed or projected from a screen	Presentation attack
Screen detection	Detecting the presence of glare or reflections that can be signs that a screen is being used to display an image or video to the camera	Presentation attack
Emulator detection	Detecting virtual cameras and also virtual mobile devices being run on larger machines	Injection attack

Metadata inspection	Detecting deviation from the vendor's expectation for a device's camera in terms of, for example, image size and resolution. An image or video that does not conform to the expectation may have been injected.	Injection attack
Telltale signatures	Adding watermarks to the images or videos being taken via their API or SDK. Any image or video that does not contain these telltale signatures may have been injected into the workflow.	Injection attack
API/SDK payload integrity	Cryptographic signing of the payload (containing the image or video) being sent from the vendor's API/SDK to its servers, which helps to prevent injection attacks at the network level as opposed to the device level	Injection attack
Data affirmation	Extracting data from the identity document and then checking it against another source such as an identity graph, a credit bureau or a government issuing authority	Both presentation and injection attacks
Human analyst	Using human analysts for image inspection in case ML/AI model does not generate a high confidence score	Both presentation and injection attacks

Source: Gartner (January 2024)