Cybercrime Atlas

Hosted at the **World Economic Forum**

Unmasking Cybercrime: Strengthening Digital Identity Verification against Deepfakes

JANUARY 2026

Images: Adobe Stock

Contents

Executive summary	3
01 Background and motivation	4
02 How the attack works: methodology and TTPs	6
03 Evaluation of deepfake tools	8
04 Threat forecasting	14
05 Recommendations and countermeasures	16
Conclusion	19
Appendix: Probability yardstick	20
Contributors	21
Endnotes	22

Executive summary

Deepfakes mean identity has become synthetic, scalable and weaponizable.

Deepfakes – artificial intelligence (AI)-generated audio and visual media that convincingly imitate real people – have rapidly evolved from entertainment tools into a material threat to digital identity systems.[1] Their misuse in know-your-customer (KYC) and remote verification processes now creates financial, operational and systemic risks for any institution that relies on digital trust.

Face-swapping attacks already span three levels:

- **Individual**: Fraudsters can open accounts, take out loans or conduct transactions using synthetic identities, while manipulated media can be used to damage reputations.
- **Organizational**: Attackers can bypass onboarding and KYC controls, impersonate staff or executives, steal data and trigger high-value kinds of fraud (such as unauthorized wire transfers).
- **Systemic**: At scale, these attacks erode confidence in digital commerce, weaken regulatory compliance and threaten the stability of broader financial ecosystems.

An analysis of 17 face-swapping tools and related camera injection techniques confirms a clear shift: while many tools remain imperfect, a subset already deliver real-time, high-fidelity impersonation capable of undermining digital KYC. Threat actors increasingly combine stolen or AI-generated identity documents, high-quality face-swap media and camera injection methods to defeat live verification.

Over the next 12–15 months, five trends will accelerate risk: widespread access to advanced AI tools, increased targeting of financial services and cryptocurrency, higher-fidelity face swaps, growth of scalable injection attacks and fragmented global regulation.

This paper outlines concrete recommendations for three key stakeholder groups:

- **KYC providers**: Invest in stronger liveness and injection attack detection, synthetic media forensics and real-time anomaly monitoring.
- **Fraud and risk teams**: Shift to risk-based monitoring that correlates identity signals across channels, incorporate threat intelligence feeds on emerging deepfake tooling, and regularly stress-test verification pipelines.
- **Financial institutions**: Establish governance frameworks that mandate resilience testing, ensure procurement standards reflect modern AI-driven threats and coordinate with regulators to accelerate convergence on deepfake-aware controls.

Deepfakes mark a turning point in cybercrime: identity itself has become synthetic, scalable and weaponizable. Sustaining trust in digital identity systems will require coordinated action, innovation and a shared commitment to security standards. The institutions that adapt early will be best-positioned to protect customers, safeguard digital ecosystems and preserve the integrity of global financial infrastructure.

01

Background and motivation

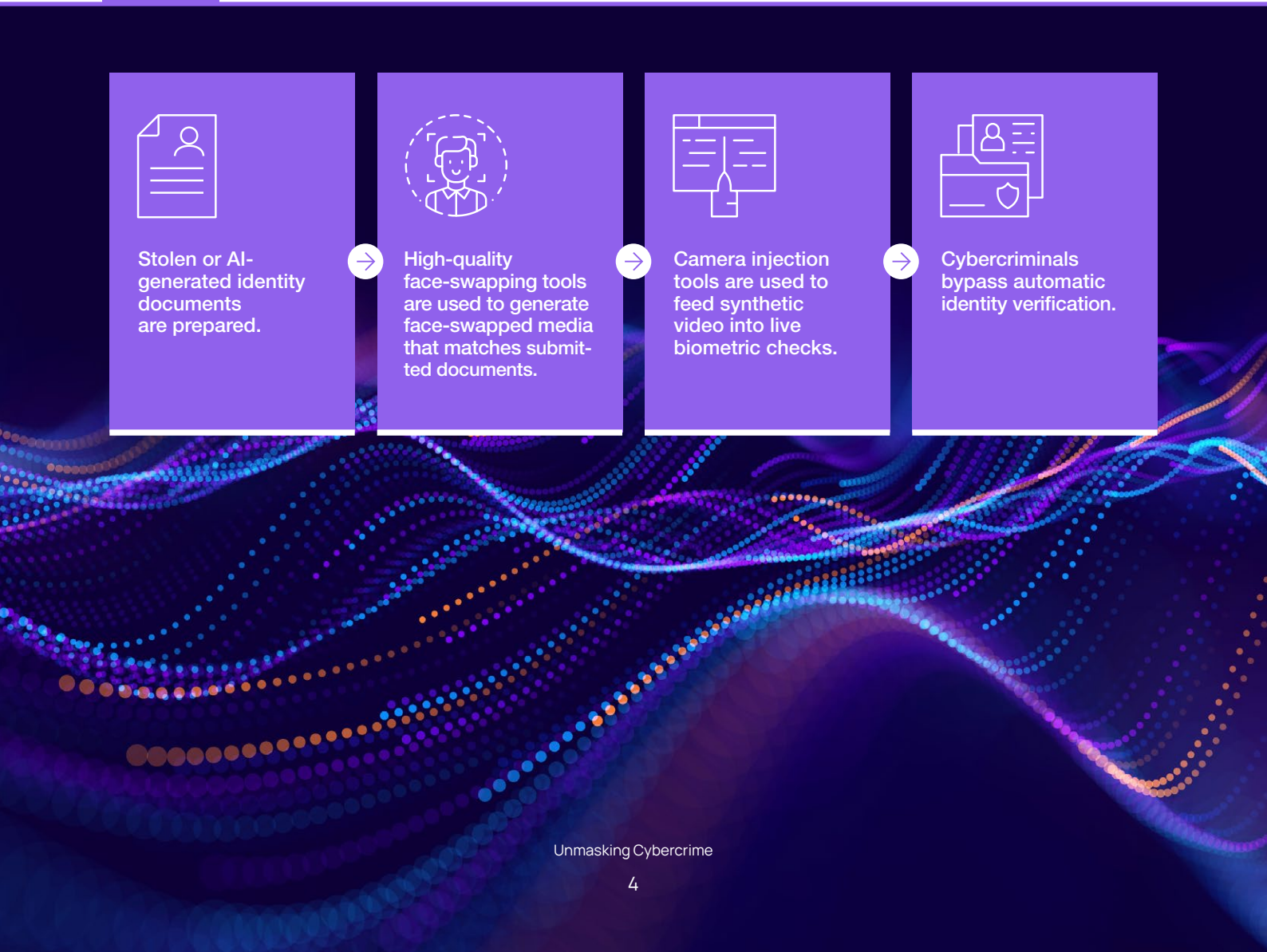
Face swapping is an AI-enabled form of impersonation and identity theft.

What are deepfakes and face swapping?

The term deepfake refers to synthetic audio or visual media that convincingly imitates real people, created through machine learning methods (typically deep neural networks).[2] This technology enables the creation of highly realistic but fabricated images, videos and audio recordings that are often hard to distinguish from authentic content.[3] Originally developed for entertainment and creative applications, deepfake technology has increasingly been exploited for malicious purposes, including fraud, disinformation and impersonation.[4]

One prominent technique within the domain of deepfakes is **face swapping**.[5] Face swapping involves replacing the facial identity of an individual in an image or video with that of another person, while preserving expressions, movements and contextual realism.[6] When misused, face swapping can serve as a sophisticated mechanism for identity theft, enabling attackers to engage in unauthorized activities under the guise of a victim's likeness.[7]

FIGURE 1 **Anatomy of a deepfake attack**



The risks posed by face swapping

The risks arising from face swapping span multiple levels – individual, organizational and systemic – introducing the risk of both financial[8] and reputational harm.

- **Individual level**: Personal identities can be exploited through face-swapped media for fraudulent activities such as unauthorized account openings, illicit loan applications and non-consensual financial transactions.[9] Furthermore, manipulated media can be weaponized to damage reputations by depicting individuals engaging in activities they did not undertake.
- **Organizational level**: Corporations face heightened risks from face-swapping attacks, particularly in processes dependent on digital identity verification, such as recruitment and remote onboarding. Attackers may bypass know-your-customer (KYC) procedures, impersonate employees or executives to exfiltrate sensitive data, or deceive individuals into authorizing fraudulent wire transfers. The credibility of business operations and the reliability of digital infrastructure are directly threatened by such attacks.
- **Systemic and societal level**: At broader scales, as trust forms the cornerstone of digital commerce and regulatory compliance, the proliferation of face-swapping attacks could undermine trust in digital ecosystems, erode public confidence in online transactions and enable large-scale data breaches, destabilizing foundational systems on which economies and societies rely.

KYC and its inherent risks

KYC verification is a critical process that organizations use to authenticate the identity of new customers and assess potential risks associated with them. KYC implementations commonly combine at least:

- **Document verification**: collection and automated validation of government-issued identity documents (passport, ID card, driver's licence)
- **Biometric verification**: comparison of a live biometric sample (commonly a facial image or short video) against the identity document

Nowadays, many organizations offer digital KYC processes, allowing customers to authenticate their identity online without physical presence. While digital KYC brings convenience, it also presents hidden risks[10] as it creates more opportunities for attackers to enter an organization's systems, particularly when identity verification relies solely on automated or unsupervised checks.[11]

Purpose and scope of this research

The main purpose of this research is to examine face-swapping attacks targeting KYC processes that rely on face verification systems. The study comprehensively evaluated the tactics, techniques and procedures (TTPs) employed by threat actors, as well as commonly used face-swapping tools. Looking beyond the present threat environment, this research provides a forward-looking threat analysis to predict potential developments in face-swapping attacks over the next one to three years. It offers recommendations to relevant stakeholders in strengthening defences against the misuse of face-swapping technologies. Additionally, the paper includes a set of actionable recommendations that organizations should consider and implement to safeguard against face-swapping attacks targeting KYC verification processes.

02

How the attack works: methodology and TTPs

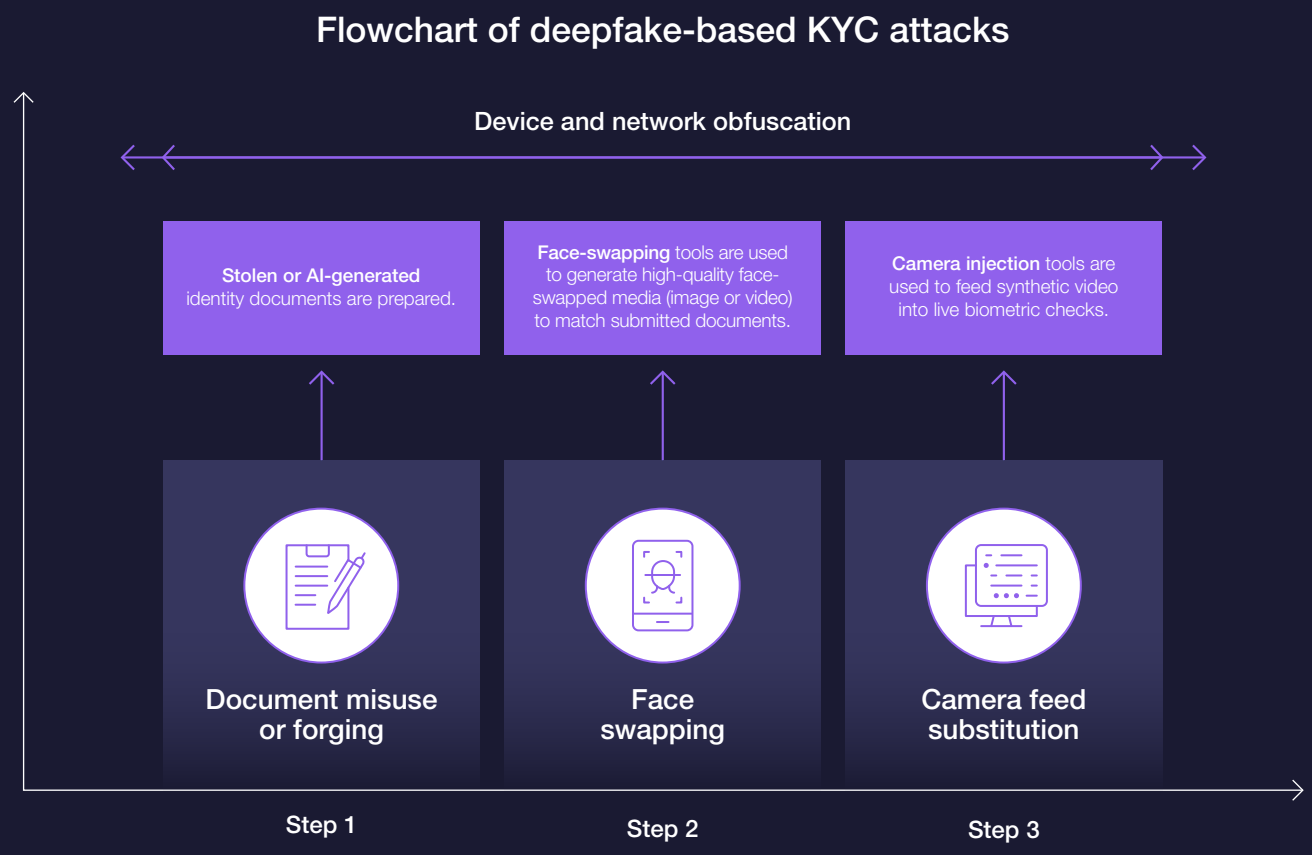
Deepfakes facilitate theft, money-laundering and the creation of fake financial services accounts.

Objective and flowchart of deepfake-based KYC attacks

The principal objectives of deepfake-facilitated KYC attacks are to onboard accounts that are not legitimately controlled by the applicant for malicious purposes such as loan fraud, money laundering, payout collection or other forms of financial abuse. Their workflow typically encompasses: 1) stolen or AI-generated identity documents; 2) high-quality face swap media to match submitted documents; 3) camera injection tools that feed synthetic video into live biometric checks.[12]

The combined effect is scalable account creation that bypasses both document and biometric KYC controls and weakens device-based signals used for fraud detection.[13]

FIGURE 2 **Deepfake-based KYC attack**



Source Cybercrime Atlas.

Detailed multi-stage attack flow

A multi-stage flow is commonly followed when face swapping is used to bypass KYC:

1. **Preparation and reconnaissance**: Targeting criteria are defined on either bulk account creation or targeted impersonation.

2. **Document misuse or forging**: A document is either stolen, purchased, forged or AI-generated to match the target identity.

3. **Face swapping and synthetic media**: A face swap model or pipeline is trained or configured to create synthetic images or videos that match identity documents. These tools may provide both pre-generated and real-time synthetic media.[14]

4. **Camera feed substitution**: Camera injection tools are used to replace a device's default live camera feed with AI synthetic media. When the KYC system captures the "live" biometric, it receives the forged media instead of the actual user, allowing the attacker to pass biometric checks.

5. **Device manipulation**: Emulators, virtual environments or other device manipulation techniques are rotated or re-provisioned to avoid linking multiple account registrations to a single physical resource.

6. **Network layering**: Residential proxies, mobile proxies or VPNs (virtual private networks) are used to diversify IP (internet protocol) addresses and geolocation signals, and to defeat simple IP-based correlation.

03

Evaluation of deepfake tools

Legitimate content creation tools are being adapted for criminal purposes.

Overview of evaluation process

Numerous face-swapping methods and camera injection tools are discussed or advertised on the internet, with some claiming the capability to bypass digital KYC systems. To assess whether these claims are well-founded, and to characterize the current deepfake landscape, a set of widely discussed face-swapping and camera injection tools was collected from darknet forums, Telegram channels and other social media platforms for the period 1 July 2024 to 30 April 2025.

The evaluation process followed the criteria listed in Table 1 and was applied separately to the face-swapping applications and camera injection toolsets. The evaluation covered 17 face-swapping tools and eight camera injection tools. To reduce the potential for misuse, vendor identities, discovery methods and step-by-step exploitation techniques have been redacted from this paper. High-level technical observations, aggregated metrics and non-exploitable examples are presented.

TABLE 1 **Evaluation criteria**

Category	Evaluation criteria
1 Face-swapping tools (fake media generators)	
Tool basics	Tool name, version, source (GitHub, Telegram, .onion), price/licensing
Media type	Static image swap/real-time video/3D avatar/animation
Real-time capability	Can it operate in live mode (e.g. front cam selfie KYC)?
Facial expression sync	Smile, blink, head turn, eye tracking supported?
Lighting adaptation	Does the fake face adapt to real lighting changes?
Voice integration	Built-in voice clone? Compatible with external voice clone tools?
Target use cases	KYC, adult content, livestreaming, impersonation
Quality metrics	Artefact presence (e.g. flicker, blending), resolution, smoothness
Model type	Generative adversarial network (GAN)-based, encoder-decoder, neural radiance fields (NeRF), StyleGAN, Avatarify, etc.
Device compatibility	Android, iOS, PC; rooted/jailbroken required?
Resource demands	Graphics processing unit (GPU)/central processing unit (CPU)/memory use during real-time rendering
Spoof detection bypass	Can it fool blink/liveness/selfie video detection?
Output format	Export as video/image or directly stream to camera (VCam)?
Toolchain dependencies	FFmpeg, TensorFlow, OpenCV, face landmarks library, etc.
Network behaviour	Connects to remote server? Downloads model weights?
Customization	Can the fraudster upload and replace specific face ID (victim photo/video)?

TABLE 1 **Evaluation criteria** (continued)

Category	Evaluation criteria
2 Camera injection tools (camera feed replacement/spoofing)	
Tool basics	Tool name, version, source, price/licensing
Injection method	Virtual camera, hooking camera application programming interface (API), overlay, emulation
Mobile platform support	Android/iOS/Emulator (BlueStacks, LDPlayer, etc.)
Root/jailbreak required	Yes/no; if yes, specify framework (Magisk, Xposed, etc.)
Camera replacement type	Full feed replacement/overlay/mirrored injection
Live feed streaming	Can it stream face-swapped content in real time?
Compatibility with KYC apps	FaceTec, Onfido, Jumio, etc. – which apps can it spoof?
Detection resistance	Can it bypass software development kits (SDKs), detecting VCam or hook tools?
Timing accuracy	Is the feed synced with user interface (UI) prompts (e.g. smile now, blink now)?
UI hijack capabilities	Can it fake app UI or overdraw the KYC app interface?
Latency and quality	Frame rate, feed lag, resolution distortion
Device artefacts	Custom package name, process ID, camera access logs
Sandbox evasion	Can it avoid being detected in test/sandboxed KYC environments?
Network behaviour	Does it connect to command-and-control (C2) or upload feeds/logs externally?
Persistence	Can it autostart, run in stealth mode or survive reboot?

Evaluation of face-swapping tools

Summary

In total, 17 face-swapping tools were evaluated to determine their functional characteristics and potential to enable digital KYC bypass. Tools were observed to cluster into three operational modes: real-time webcam swappers, offline desktop frameworks and hosted web services. Genuine real-time face swapping capability was uncommon (demonstrated by five tools), and only a subset of those (three tools) provided virtual-camera or equivalent injection paths suitable for integration into live verification flows. Overall, the greatest KYC risk was found where low-latency, high-fidelity, real-time swaps were deliverable directly into a verification pipeline. Most tools were intended for creative or entertainment use and did not explicitly include anti-KYC functionality.

These findings are based solely on claims made in the tools' publicly available documentation and websites, not on direct empirical testing.

Operational modes and real-time capability

Tools were found to operate in three principal modes:

- **Real-time webcam swappers** – capable of low-latency manipulation and virtual camera output; out of five tools demonstrating genuine real-time face swapping, three enabling virtual-camera or equivalent injection paths that could feed a verification flow with minimal intermediary processing
- **Offline desktop frameworks** – designed for post-production workflows and high-quality output, but without instantaneous rendering
- **Hosted web services** – browser-accessible services that performed server-side processing (some offering "real-time preview" but requiring final cloud processing, limiting continuous live use)

Real-time capability was concentrated in a small number of technically mature platforms, according to research and open-source information collected on these tools rather than through direct testing. Where real-time output was unavailable, hosted or offline outputs could still be chained with injection techniques to simulate live inputs, albeit with additional detectable signals.

Expression and motion fidelity

Tracking of dynamic facial features – blink rate, lip motion, head pose and micro-expressions – was identified as a critical determinant of whether outputs could pass liveness checks. High-quality dynamic expression preservation was uncommon; six out of 17 tools employed advanced motion models or architectural approaches that preserved nuanced expressions, producing more natural output for dynamic video. The majority of tools provided partial or static expression reproduction, which increased susceptibility to detection under challenge–response or randomized motion prompts.

Lighting adaptation and blending quality

Advanced, environment-responsive lighting adaptation was rarely observed. Only two tools incorporated enhanced blending techniques or optimizations targeted at poor lighting conditions. These two were effective primarily on static or pre-recorded content and required manual tuning for best results. Common failure modes included colour tone mismatch, shadow inconsistency and visible edge seams that were amplified by compression and abrupt changes in illumination. When present, lighting adaptation typically focused on skin tone matching and basic face blending rather than real-time adaptive relighting.

Latency, timing accuracy and robustness to challenge–response

Latency and timing accuracy were key limiting factors for passing real-time liveness checks. Timing slippage during challenge–response interactions was consistently observed as a failure point. Tools that produced low-latency output with accurate synchronization between facial motion and audio or prompts were rare. Many tools advertised "real-time preview" capabilities, yet required additional cloud processing for final output, resulting in latency unsuitable for continuous live impersonation.

Audio integration and voice cloning

Native voice cloning functionality was not provided by any of the evaluated tools. Overall, three tools offered basic audio features such as pre-generated or AI-synthesized speech insertion or lip-sync to supplied audio tracks. These features did not equate to full voice replication from sample audio. Therefore, creation of convincing audio-visual deepfakes would generally require integration with an external, dedicated voice synthesis platform.

Platform support, device compatibility and resource demands

Platform support was distributed as follows: three tools explicitly supported PC (desktop) environments, four were compatible with Android, and two supported iOS. Browser-based implementations were common (10 tools), reflecting a trend towards accessibility and ease of use. Cross-platform support was moderate, with most tools supporting approximately two platforms on average. Overall, nine tools required local installation and execution on user devices, while eight operated entirely online with server-side processing. Desktop real-time tools were observed to require capable graphics processing units (GPUs) for stable, low-latency operation; hosted services reduced local hardware requirements at the cost of network dependency.

Network behaviour

Hosted solutions were noted to be network-dependent; observed behaviours included sustained upstream video upload and frequent small data exchanges for incremental processing. Tools that operated locally produced minimal network traffic, but injection techniques that forwarded altered frames into remote verification flows introduced identifiable network patterns (for example, consistent outbound streams followed by virtual camera input).

Typical use cases and intent

Most tools (approximately 11) were positioned towards creative, entertainment and social media use cases (memes, parody, live streaming, avatar creation). In total, four tools were marketed for legitimate professional applications – digital marketing, visual effects (VFX), academic research and artist assistance. Meanwhile, two tools appeared to be built for security testing or red team operations. Despite stated benign intentions for some platforms, capabilities such as high-quality video generation and real-time swapping present potential for misuse in identity spoofing and KYC bypass scenarios.

Observed artefacts and indicators for detection

Consistent artefact classes that may be harnessed for detection were identified:

- **Temporal artefacts**: timing slippage during challenge–response, irregular blink/lip synchronization
- **Spatial seams**: edge blending inconsistencies, flicker amplified by compression
- **Lighting mismatches**: abrupt colour or shadow discontinuities during illumination changes
- **Motion inconsistency**: loss of micro-expressions and unnatural head pose transitions
- **Network patterns**: atypical combinations of upstream video traffic with virtual camera injection

These artefacts were more pronounced in tools without advanced motion or lighting models, and were often exacerbated by low bandwidth, device resource constraints or aggressive compression in verification pipelines.

KYC risk implications

Risk was found to be highest where the following conditions were jointly met: 1) genuine real-time swapping capability, 2) virtual camera or injection path into the verification flow, and 3) high dynamic expression fidelity and reasonable lighting blending. Tools that were offline or cloud-based without virtual camera injection posed lower immediate risk to live liveness checks, but could still be repurposed when combined with injection techniques or used in conjunction with separate voice synthesis tools. Image-only manipulations presented the lowest immediate risk to active liveness mechanisms unless coupled with additional tooling.

Evaluation of camera injection tools

Summary

In total, eight camera injection tools were evaluated to determine their technical behaviours, platform coverage, persistence characteristics and potential to facilitate digital KYC bypass. The sample comprised a mix of free and commercial solutions; three tools were freely accessible, and five were commercial (price range: approximately $10–3,000). Overall, three tools supported both pre-recorded and livestreaming workflows. Another three tools supported only pre-prepared media. Tools were observed across Android, iOS and Windows platforms. Overall, camera injection capabilities were found to be technically diverse but constrained by latency, content format requirements and detectable device artefacts. As a result, most tools seem limited in their ability to reliably defeat modern KYC systems that use dynamic prompts and software development kit (SDK)-level integrity checks.

Limitations of the evaluation

The assessment was limited to high-level observations and aggregated testing. Full validation against every commercial KYC product was not performed. Tool capabilities may have changed following the collection window; vendor updates or new driver signatures may alter the detection surface. Additionally, the analysis covered only a small subset of deepfake tools available in the wild, and the evaluated samples may not fully represent the diversity or sophistication of the broader ecosystem. Sample bias was possible due to the sources used to obtain these tools.

The scope of the paper encompasses open-source intelligence (OSINT) findings and case analyses collected between early 2024 and late 2025. While the paper is based on current technological capabilities and observed threat activity, it should be noted that the rapid pace of AI development may render some conclusions less applicable in the future.

Injection method classification

Three injection approaches were observed and used to classify the evaluated tools:

- **App-level injection**: Four tools performed injection by intercepting camera application programming interface (API) calls (e.g. Camera2 on Android, AVFoundation on iOS) and supplying synthetic content (pre-recorded video or manipulated stream) in place of the physical camera feed. This approach commonly relied on root/jailbreak and frameworks like Magisk/Xposed (Android) or Substrate (iOS). These are easier to implement compared to other injection methods, and usually detectable with root/jailbreak check or hooking frameworks like Xposed/Substrate.

- **System-level virtual camera drivers**: Three tools installed drivers or virtual camera devices at the OS level (for example, a Windows virtual webcam driver). These drivers appear as a physical camera to all apps (e.g. Zoom, Teams, KYC software), making detection by app-level checks more difficult. The injected content can be a live deepfake or pre-recorded video. Since a system-recognized hardware driver or any app that uses the OS webcam sees this as a legitimate device, it is hard to detect this level of injection.

- **Hybrid overlay/mirroring**: One tool combined container/virtualization techniques with user interface (UI) overlays or stream mirroring. Overlay methods rendered synthetic content on top of the real camera preview at the UI level, while mirroring techniques duplicated and redistributed camera streams (for instance, via DirectShow) to multiple fake camera endpoints. This approach could bypass certain app checks but could leave artefacts (e.g. duplicate camera feeds, mirrored sessions).

Platform support, price and live feed streaming

Pricing models vary widely among the tools. There are two open-source options that are available under permissive licences – the Massachusetts Institute of Technology (MIT) and the GNU General Public Licence (GPL). Meanwhile, others follow commercial models – ranging from **affordable one-time fees** (e.g. approximately $25.95) to **premium-grade offerings** (priced at **$3,000**).

In terms of **platform support**, mobile-focused tools generally target **Android 5+**, with some requiring higher versions such as Android 9+ and **iOS 11–13.3**. Desktop tools support a wide range of **Windows OS versions (XP to 11)**.

For **live feed streaming**, only three of eight tools support **real-time messaging protocol (RTMP) or stream duplication**, while others rely on **pre-recorded or static media** without native streaming support. Orientation mismatches, resolution requirements or manual intervention may still impact reliability during streaming attempts.

Root/jailbreak requirements

Half of the evaluated tools (four) required elevated privileges (typically achieved via root or jailbreak). Commonly harnessed frameworks included Magisk, Xposed/LSPosed (Android) and Cydia Substrate (iOS). The remaining tools claimed rootless operation; however, some rootless approaches (for example, virtualization engines or containers) were found to leave detectable system footprints despite not requiring formal root access.

Latency, timing accuracy and quality

Timing accuracy: Most tools struggled to maintain accurate timing during challenge–response interactions or randomized liveness prompts. Significant synchronization delays were observed for RTMP-based and static-content tools, reducing their ability to pass dynamic verification. Only one tool demonstrated real-time feed splitting with partial success in synchronized scenarios.

Video quality: Output quality varied widely and was highly sensitive to source media format, device capability and manual configuration. Orientation mismatches, resolution constraints and rendering glitches were common failure points on lower-end hardware. Optimal performance required modern central processing units (CPUs)/GPUs and correctly formatted source media.

UI hijack capabilities and networking behaviour

In total, seven evaluated camera injection tools lack UI hijack capabilities, meaning they do not overlay, manipulate or hook into application interfaces, or use AccessibilityService (an Android feature designed to enhance the UI and assist users with disabilities or those who might temporarily be unable to fully interact with their device) or system-level UI overlays. Their functionality is strictly limited to camera feed emulation or redirection.

In terms of network behaviour, all tools operate locally with no evidence of command-and-control (C2) communication, data exfiltration or external server beacons. This suggests that, from a networking and UI manipulation standpoint, these tools pose minimal active threat beyond the camera injection vector itself and also limit their functionality for remote-controlled attacks.

Persistence, stealth and operational traces

Persistence: Desktop tools commonly persisted via Windows services, scheduled tasks or installed drivers. Mobile persistence frequently relied on combination with root frameworks (e.g. modules that reload on boot). Some virtualization/container approaches remained active only while the virtual environment was installed and required explicit user action to relaunch.

Stealth: Most tools lacked strong stealth and were visible in task managers, package lists or driver registries. Root/jailbreak indicators and module identifiers were consistently recoverable.

KYC compatibility and KYC risk implications

Browser flows versus SDKs: Virtualization and feed redirection tools were observed to sometimes bypass browser-based flows (particularly when manual setup for browser permissions was performed), but these approaches generally failed against modern SDK-based verification systems that embed integrity checks and use challenge–response mechanisms.

Risk to live KYC flows was assessed as conditional:

- **Higher risk** was observed where tools combined 1) virtual camera drivers or reliable injection paths, 2) real-time streaming capability, and 3) low latency with accurate timing.
- **Moderate to low risk** was attributed to rootless or offline tools that relied on pre-recorded media, strict format requirements or manual setup. These tools could still be repurposed for replay or injection attacks, but the additional steps increased the number of detectable signals.

Detection and risk mitigation

Detection and risk mitigation efforts were advised to prioritize the following signals and controls:

1. **Hooking framework detection**: Flag the presence of Xposed/LSPosed/Magisk modules and other hooking artefacts.
2. **Virtual device enumeration**: Monitor for unusual or duplicate camera device registrations and non-standard device class IDs.
3. **Filesystem indicators**: Scan for non-standard media directories, override files, cloned package names and container-related package families (e.g. com.fvbox.*).
4. **Driver and registry monitoring (desktop)**: Detect installation of signed virtual webcam drivers, recently added registry keys associated with device drivers, and unexpected Windows services.
5. **Timing and challenge correlation**: Implement strict challenge–response timing checks and multi-modal synchronization (audio versus visual) to increase false positive resilience and detect timing slippage.
6. **Network and process telemetry**: Correlate local upstream streaming patterns with the creation of virtual camera devices to detect feed injection attempts.
7. **SDK integrity**: Where feasible, prefer SDK-based verification modules with embedded integrity checks over pure browser flows.

04

Threat forecasting

The effectiveness of threats to digital KYC is evolving at speed.

Summary

The evolution of face swap threats to digital KYC is shaped by rapid generative AI (genAI) development, fragmented regulation and the adaptability of criminal ecosystems.[15] While the currently evaluated tools show limited ability to bypass advanced KYC systems, rapid improvements in model realism and accessibility indicate that this barrier may soon diminish.[16]

Over the next 12–15 months, five trajectories are assessed as most likely:

1. Democratization of AI tools lowering entry barriers and increasing attack complexity
2. Persistence of finance and cryptocurrency as prime targets, with expansion into other KYC-dependent sectors
3. Rising fidelity of face-swap technology enhancing realism and undermining verification
4. Persistence of presentation attacks in the near term, with injection attacks escalating as active liveness adoption grows
5. Fragmented regulation constraining defences in the short term, but regulatory convergence likely improving resilience in the medium term

Forecast 1
Democratization of AI tools and expansion of attack complexity

The short-term trajectory of face swap threats is highly likely directed towards greater scale, fuelled by the democratization of genAI. Open-source repositories and underground forums already distribute deepfake toolkits tailored for KYC bypass, and industry assessments (e.g. iProov, Trend Micro , Group-IB) confirm the growth of such offerings. Meanwhile, academic platforms like **Civitai** report millions of model downloads. These developments reduce the time and expertise required to create synthetic identities (sometimes achievable within hours using consumer hardware).

As accessibility increases, the complexity of attacks is also expected to grow. Multiple injection and presentation techniques can be combined in layered approaches, expanding the attack surface for defenders.

Forecast 2
Finance and cryptocurrency as prime targets, expansion to other sectors

Financial services and cryptocurrency exchanges remain the most attractive targets due to their high monetary value, reliance on remote biometric onboarding and uneven regulatory standards. Attacks in these sectors support fraud, money laundering and unauthorized account creation.

Other industries adopting KYC, including gambling, telecommunications and trading, are increasingly targeted. The introduction of new regulations, such as the UK's online safety rules enacted in July 2025, has already correlated with increased searches for bypass tools, underscoring attackers' responsiveness to regulatory change.

Forecast 3
Rising fidelity of face-swap technology

Advances in academic research and the entertainment industry are rapidly improving realism in face swaps and synthetic video generation, sometimes within smartphone applications. Tools such as **DreamID** (2025) outperform prior models in identity similarity and image fidelity under complex conditions, while lightweight frameworks like **FastSwap** have demonstrated real-time, high-resolution performance. Complementary technologies such as **VASA-1** and **LivePortrait** simulate microexpressions and nuanced head movements in real time.

Given the open release of code and pretrained weights, these capabilities are likely to be repurposed for malicious use. While training remains resource-intensive, inference and deployment are increasingly democratized through repositories and cloud GPU services, lowering barriers for threat actors.

Forecast 4
Presentation attacks persist, but injection threats escalate

Injection attacks are growing rapidly. Industry reporting highlights sharp increases; iProov observed a **783% increase** in injection attacks in 2024, while Jumio noted an **88% year-on-year rise** in 2025. Injection vectors, particularly those targeting mobile web applications, are expected to escalate further as active liveness detection becomes more common.

Forecast 5
Alternative scenarios and future directions

Several uncertainties could shift these trajectories. Stricter guardrails on mainstream genAI tools may constrain democratization, limiting the assumed growth in accessibility. Shifts in attacker targeting could occur if financial services become overly risky for attackers, with healthcare or gaming emerging as alternative focuses. Advances in defensive AI could offset the fraud potential of higher-fidelity models, while slower-than-expected adoption of active liveness could prolong the dominance of presentation attacks. Accelerated regulatory convergence, on the other hand, could elevate global standards more quickly than forecast.

Note that the probability of each forecast is also conditional on several triggers. Breakthroughs in AI-driven liveness simulation, accelerated rollout of regulatory standards or widespread availability of turnkey deepfake toolkits could compress the expected timelines from 12–15 months to less than a year. In contrast, effective watermarking, improved dataset provenance or tighter access controls on genAI models could delay or limit the projected escalation of attacks.

05

Recommendations and countermeasures

Defences must be adaptive, multi-layered and continuously improved.

Summary

The rapid evolution of deepfake technology has created an escalating contest between genAI capable of producing synthetic identities and fraud detection systems designed to stop them. To maintain trust in digital identity verification, defences must be adaptive, multi-layered and continuously improved. This section presents a structured set of recommendations for three major stakeholder groups within the digital KYC ecosystem: solution providers, fraud teams and financial institutions.

An effective defence strategy must therefore combine **technical rigour, risk-based monitoring and governance discipline**. The following recommendations are organized according to the roles and responsibilities of three primary stakeholders. Each set of recommendations is designed to strengthen resistance against deepfake-based attacks (particularly face swapping and camera injection attempts), while balancing privacy, compliance and operational efficiency.

Recommendations by stakeholder group

KYC solution providers (liveness and anti-spoof vendors)

KYC vendors play a critical role in detecting manipulated visual streams before identity verification is completed. The following measures are recommended:

1. **Camera path verification** – Implement mechanisms to detect or flag virtual cameras, mid-session device swaps and new driver installations. This helps verify that the video source is native and from a trusted device path, cutting off the easiest delivery route for injected or face-swapped streams.

2. **Active and dynamic liveness checks** – Use randomized prompts and dynamic lighting variations (e.g. brief screen flashes) to introduce unpredictability that exposes synchronization errors typical of real-time face swaps. Since a moving target is hard to pre-render or synchronize perfectly, this helps surface latency and sync issues common in real-time face swap pipelines.

3. **Transport-aware stream scoring** – Score the actual stream received (post-compression) for seams, flicker, lip audio drift and texture "swim". Perform quality checks on the delivered media, not the local preview. Compression often amplifies artefacts, making spoofs easier to detect.

4. **Temporal consistency monitoring** – Track frame-to-frame stability (flicker, seam drift), lip sync offset and frame pacing jitter over short time windows. Measuring visual consistency over time helps because real-time swaps often wobble across frames, while authentic human video does not.

5. **Context telemetry APIs** – Provide encoder/codec, bitrate, resolution/aspect ratio, device/OS/browser details and mid-session changes via API. These lightweight stream and device fingerprints help expose telltale anomalies of injected or synthetic video to risk engines.

6. **Explainable detection outcomes** – Return concise reason codes (e.g. "VCam suspected," "temporal jitter high", "lip sync off") to provide human-readable explanations tied to detections. This speeds up triage and appeals, enabling consistent reviewer decisions.

7. **Environment-calibrated models** – Develop model packs optimized for real-world conditions (e.g. mobile front cameras, low-light environments, low bandwidth) to reflect actual customer scenarios. This reduces false positives and improves detection rates in challenging conditions.

8. **Policy integration hooks** – Offer flexible APIs to trigger additional verification (e.g. document checks or human reviews) based on detection outcomes.

9. **Sandbox testing frameworks** – Provide a safe test suite to A/B test prompts, thresholds and device policies. This serves as a staging lab for liveness tuning and helps teams adjust configurations and quantify trade-offs before rollout.

Fraud teams (risk engines and monitoring units)

Fraud teams are responsible for operational monitoring and analytics-based risk assessment. The following practices enhance detection depth and accuracy:

1. **Trusted camera source control** – Allowlist native device cameras, and log or block sessions initiated from virtual or swapped sources. This policy and telemetry ensure trusted capture paths and prevent synthetic feeds from entering undetected.

2. **Timing correlation and latency analysis** – Record prompt timestamps and measure user reaction latency to detect non-human response patterns.

3. **Contextual signal correlation** – Gather device, browser and encoder metadata to identify anomalies linked to synthetic or automated environments.

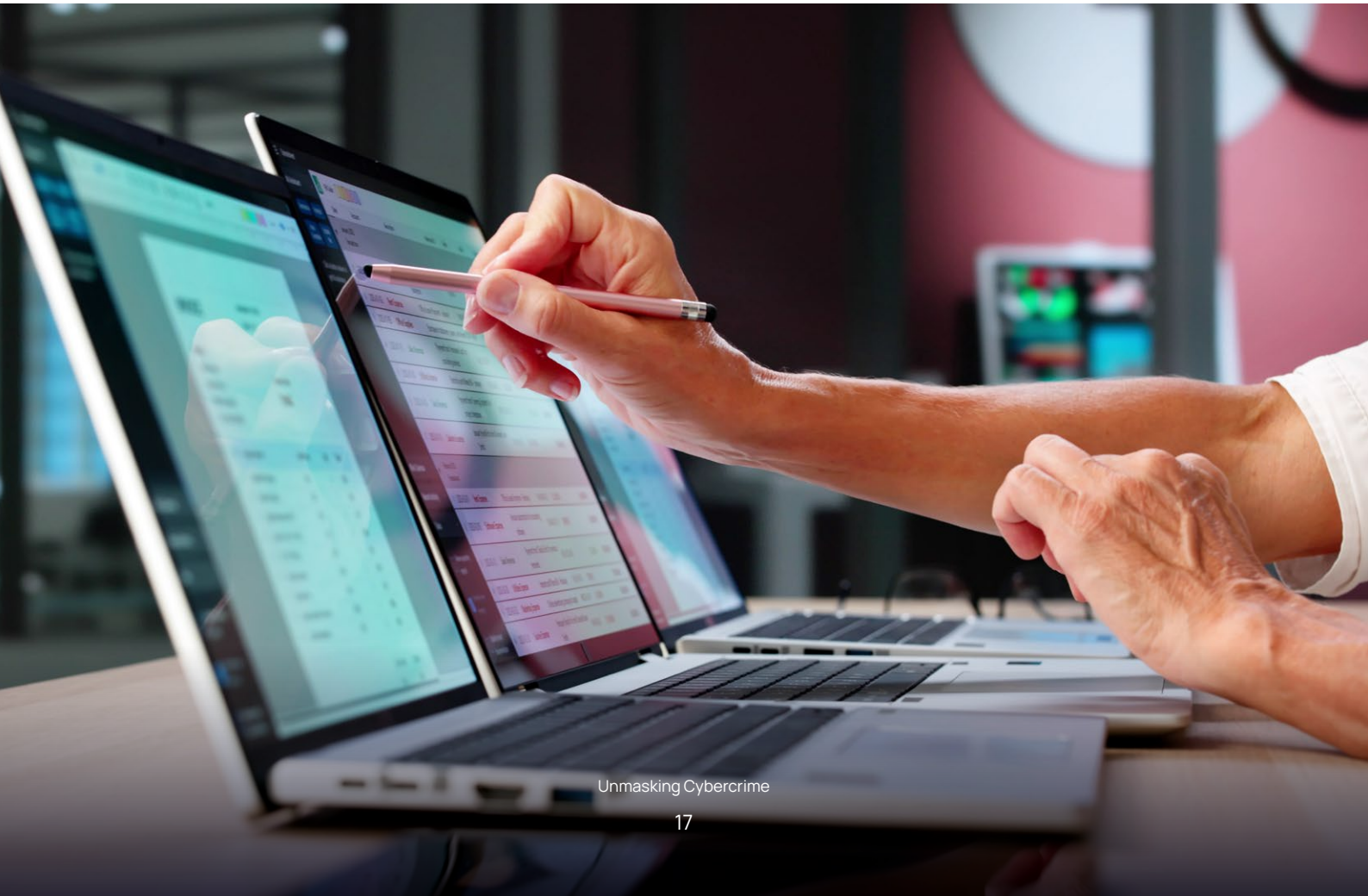
4. **Step-up verification frameworks** – Define pre-approved escalation actions when risk thresholds are exceeded (e.g. additional document checks or human reviews). This converts risk signals into controlled friction only when needed.

5. **Post-compression artefact analysis** – Inspect the ingested video stream for compression-level artefacts indicative of manipulation.

6. **Standardized case taxonomy** – Establish consistent labelling (e.g. “suspected face swap,” “timing anomaly”) to enable model feedback loops and analytical consistency.

7. **Threat chain correlation** – Combine camera anomalies, timing data and transaction risk metrics to identify multi-stage attack sequences. This helps uncover combined attacks that single checks might miss.

8. **Closed-loop feedback to vendors** – Regularly provide verified outcomes to KYC vendors to improve model performance and reduce false detections.



Institutions (programme and governance layer)

Institutions govern the broader risk management framework and ensure vendor compliance. The following governance-oriented recommendations are advised:

1. **Risk-based deployment** – Prioritize control rollouts for high-risk use cases such as account onboarding, account recovery or high-value transactions. Phased deployment based on exposure helps deliver rapid risk reduction with minimal customer impact.

2. **Privacy and retention controls** – Enforce minimal data retention, limited access and comprehensive audit trails for biometric and liveness data. These guardrails for responsible use reduce privacy and compliance risks while maintaining detection capabilities.

3. **Customer transparency** – Provide simple explanations about liveness prompts and lighting effects to reduce confusion and improve user experience.

4. **Vendor procurement standards** – Require vendors to deliver features such as virtual camera detection, compressed-stream scoring and explainable outcomes as part of procurement contracts.

5. **Governance documentation** – Maintain clear operational guidelines, thresholds and appeal mechanisms. This documentation and these operational guardrails ensure that decisions remain consistent, auditable and defensible.

6. **Staff training and playbooks** – Implement concise operational guides for customer support and escalation handling to ensure consistent responses and reduce errors.

7. **Red team testing cadence** – Conduct periodic simulation exercises to benchmark defences against evolving real-time face swap tools.

8. **Layered verification models** – Combine multiple verification signals (e.g. liveness, document validation, behavioural biometrics) to create a defence-in-depth architecture.

9. **Government-condoned intelligence sharing** – In the interest of public safety and national resilience, companies are permitted to share intelligence on tools or related information with government agencies and peer organizations (provided such disclosures are proportionate, purpose-specific and aligned with lawful exemptions to privacy legislation).

10. **Wider implementation of government legislation to protect "right to identity"** – Laws granting individuals copyright over their own face, voice and body could strengthen protections against face-swapping technologies and other deepfake threats, such as Denmark's proposed "right to identity" law.

Conclusion

The research demonstrates that deepfake technology has transitioned from a niche novelty into a **scalable tool for organized fraud**, capable of threatening the integrity of digital KYC systems. While current deepfake tools vary widely in sophistication, the trend points towards increasing accessibility, automation and realism – making the threat both persistent and adaptive.

The evaluation of 25 tools across different platforms has shown that even moderate-quality face swapping models, when integrated with camera injection techniques, can deceive certain biometric systems under specific environmental or technical conditions. Most attacks, however, still exhibit detectable inconsistencies, particularly in temporal synchronization, lighting and compression artefacts. These weaknesses provide actionable focus points for advanced detection models and forensic countermeasures.

The study also reveals that the **defensive landscape must evolve in tandem** with genAI advancements. Detection models must not only recognize known patterns but anticipate future ones through continual learning, feedback integration and cross-platform signal correlation. As adversaries harness open-source AI models and low-cost hardware, the barriers to executing real-time identity spoofing will continue to decline, demanding equally agile defences.

The research group outlined key areas for future work:

- Assessing deepfake KYC bypass security across operating systems (iOS versus Android)
- Evaluating the performance of different genAI models
- Reviewing data on processing speed, memory use, input triggers and noise levels
- Considering the effects of new regulations on deepfakes and face swapping
- Investigating social engineering risks linked to face swapping

Moving forward, resilience against deepfake-enabled KYC bypass will depend on the **synergistic effort of three layers**:

1. **Technological innovation** – development of transport-aware and temporally consistent anti-spoofing systems
2. **Operational vigilance** – adaptive fraud analytics and risk escalation frameworks
3. **Governance and collaboration** – unified industry standards, responsible data management, reformed government policies and red team testing practices

Ultimately, while deepfake-driven fraud may never be fully eliminated, it can be **contained, deterred and made economically unviable** through layered defences, transparency in model design and active intelligence sharing among financial institutions, vendors and regulatory stakeholders.

This paper therefore emphasizes that **deepfake threats to digital KYC represent not only a technical challenge but a systemic one** – requiring persistent innovation, collaboration and foresight to ensure the continued trustworthiness of digital identity verification in the age of synthetic reality.

Appendix:
Probability yardstick

The probability language used for the forecasts aligns with the framework adopted by the UK Professional Head of Intelligence Assessment (PHIA), which standardizes estimative terms with approximate numerical ranges.

TABLE 2 **Professional Head of Intelligence Assessment (PHIA) probability yardstick**

Associated term	Probability range
Almost certainly	Over 95%
Very likely	80–90%
Likely	65–75%
Even chance	45–55%
Unlikely	20–35%
Very unlikely	Less than 10%

Source UK Government. (2025). *Explaining Uncertainty in UK Intelligence Assessment.*

Contributors

Lead author

Natalia Umansky
Project Specialist, Cybercrime Atlas,
World Economic Forum

World Economic Forum

Seán Doyle
Lead, Cybercrime Atlas,
World Economic Forum

Research leads

Banco Santander

Group-IB

Acknowledgements

We extend our deepest thanks to the researchers, investigators, engineers and experts from the Cybercrime Atlas community who contributed to this report. For reasons of operational security, they cannot be named, but the findings shared here are based on insights developed during collaborative research by experts from the following organisations participating in the Cybercrime Atlas community:

Lemon

Mastercard

Recorded Future

SpyCloud

Trend Micro

Production

Louis Chaplin
Editor, Studio Miko

Laurence Denmark
Creative Director, Studio Miko

George Messer
Designer, Studio Miko

Cat Slaymaker
Designer, Studio Miko

Endnotes

1. Sumsub. (2023). *Sumsub Research: Global Deepfake Incidents Surge Tenfold from 2022 to 2023*. https://sumsub.com/newsroom/sumsub-research-global-deepfake-incidents-surge-tenfold-from-2022-to-2023/.

2. Unit 42. (2025). *False Face: Unit 42 Demonstrates the Alarming Ease of Synthetic Identity Creation*. https://unit42.paloaltonetworks.com/north-korean-synthetic-identity-creation/.

3. Trend Micro. (2025). *Deepfake It Till You Make It: A Comprehensive View of the New AI Criminal Toolset*. https://www.trendmicro.com/vinfo/us/security/news/cybercrime-and-digital-threats/deepfake-it-til-you-make-it-a-comprehensive-view-of-the-new-ai-criminal-toolset; Group-IB. (2024). *Deepfake Fraud: How AI is Deceiving Biometric Security in Financial Institutions*. https://www.group-ib.com/blog/deepfake-fraud/.

4. Cox, J. (2025). The Age of Realtime Deepfake Fraud Is Here. *404Media*. https://www.404media.co/the-age-of-realtime-deepfake-fraud-is-here/; TRM Labs. (2025). *Sextortion, deepfakes, and the financial infrastructure behind digital exploitation*. https://www.trmlabs.com/resources/blog/sextortion-deepfakes-and-the-financial-infrastructure-behind-digital-exploitation.

5. Sensity. (2024). *The State Of Deepfakes 2024*. https://5865987.fs1.hubspotusercontent-na1.net/hubfs/5865987/SODF%202024.pdf.

6. Semla. (2025). *iGaming fraud up 73%. Deepfake & ID fraud escalate*. https://www.getsemla.com/blog/igaming-fraud-up-73-deepfake-id-fraud-escalate.

7. Hawkins, W., Russel, C. & Mittelstadt, B. (2025). Deepfakes on Demand: the rise of accessible non-consensual deepfake image generators. *FAccT '25: Proceedings of the 2025 ACM COnference on Fairness, Accountability, and Transparency*. https://arxiv.org/html/2505.03859v1.

8. Payment Village. (n.d.). *How I used deepfakes to bypass security verifications in a bank*. https://www.paymentvillage.org/blog/how-i-used-deepfakes-to-bypass-security-verifications-in-a-bank.

9. Sift. (2025). *Digital Trust Index Q2 2025: AI Fraud Data & Insights*. https://sift.com/index-reports-ai-fraud-q2-2025/.

10. Vincent, J. (2022). Liveness tests used by banks to verify ID are 'extremely vulnerable' to deepfake attacks. *The Verge*. https://www.theverge.com/2022/5/18/23092964/deepfake-attack-facial-recognition-liveness-test-banks-sensity-report; Cox, J. (2023). How I Broke Into a Bank Account With an AI-Generated Voice. *Vice*. https://www.vice.com/en/article/how-i-broke-into-a-bank-account-with-an-ai-generated-voice/.

11. TechRadar. (2025). *Cybercriminals are deploying deepfake sentinels to test detection systems of businesses*. https://www.techradar.com/pro/security/cybercriminals-are-deploying-deepfake-sentinels-to-test-detection-systems-of-businesses-heres-what-you-need-to-know; Axonlabs. (2024). *iBeta Liveness trends H1 2024*. https://axonlabs.pro/ibeta-liveness-trends-2024/.

12. During this process, device and network obfuscation such as emulators, virtual spaces, VPN, or hosting service may be used to defeat device-linkage and volume-based detection.

13. CATO Network. (2024). *Cato CTRL Threat Research: ProKYC – Deepfake Tool for Account Fraud Attacks*. https://www.catonetworks.com/blog/prokyc-selling-deepfake-tool-for-account-fraud-attacks/.

14. Veriff. (2025). *Real-time deepfake fraud in 2025: Fighting back against AI-driven scams*. https://www.veriff.com/identity-verification/news/real-time-deepfake-fraud-in-2025-fighting-back-against-ai-driven-scams.

15. TRM Labs. (2025). *AI-Enabled Fraud: How Scammers Are Exploiting Generative AI*. https://www.trmlabs.com/resources/blog/ai-enabled-fraud-how-scammers-are-exploiting-generative-ai ; Group-IB. (2025). *What Are Deepfakes? How to Detect Them*. https://www.group-ib.com/resources/knowledge-hub/deepfakes/.

16. Carroll, M. (2025). Age verification tools on adult websites bypassed in seconds. *Sky News.* https://news.sky.com/story/hackers-prove-age-verification-systems-on-pornography-sites-can-be-bypassed-in-seconds-13401733.

Cybercrime Atlas

Hosted at the **World Economic Forum**