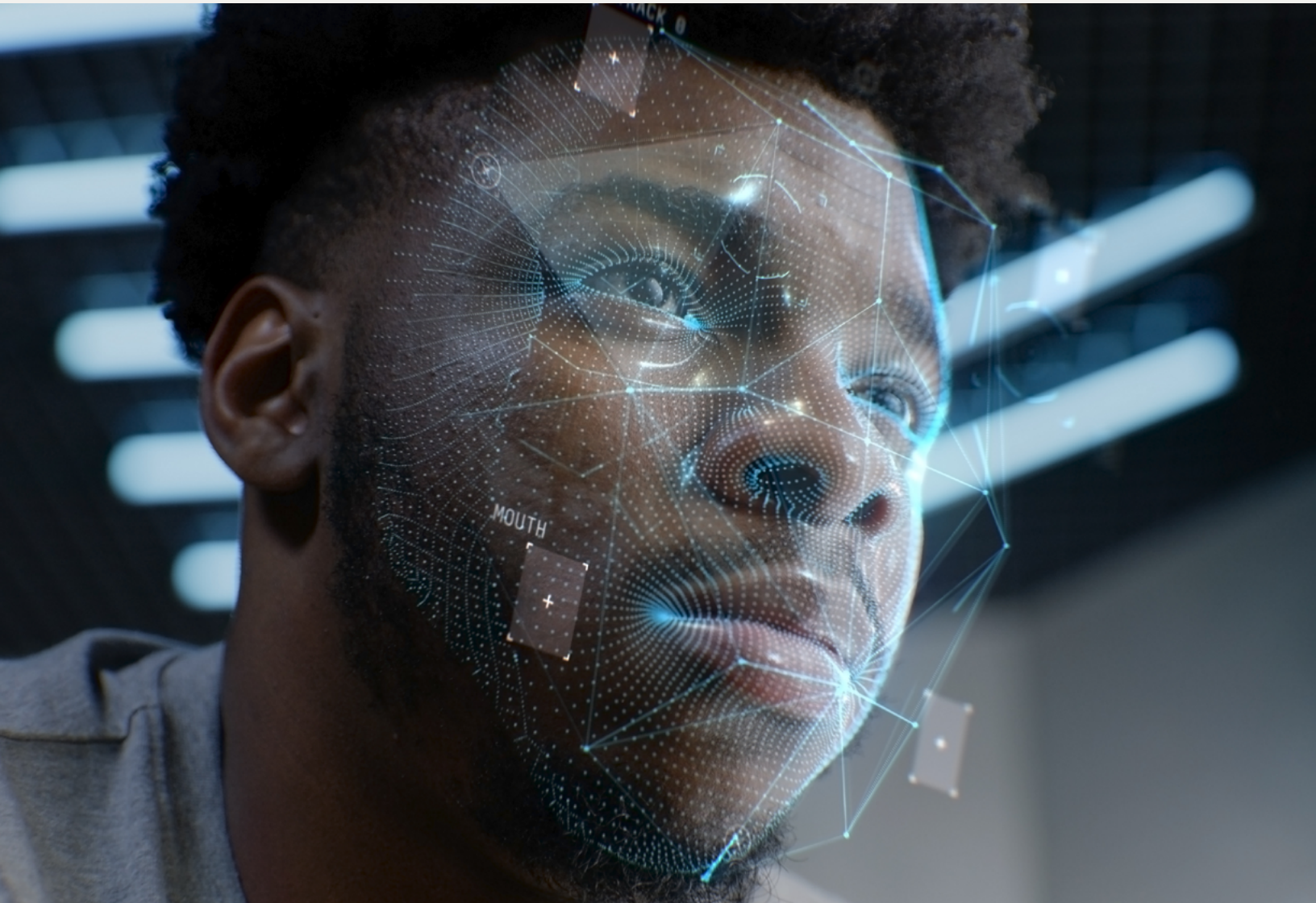
kyndryl

How to protect your enterprise from deepfakes

by Kris Lovejoy

Global Security and Resiliency Practice Leader, Kyndryl



Key takeaways

AI-powered deception is now a boardroom risk. The ability to recognize and defend against deepfake attacks will become a defining skill for enterprises.

Human trust is the new attack surface. Cybercriminals have industrialized social engineering, making active verification and multi-layered defense essential.

Private knowledge is the ultimate safeguard. When deepfakes are built from public data, verification that relies on shared, private information will help thwart attacks.

Introduction

AI-powered deception now poses a clear and growing danger to every enterprise.

In a fundamentally altered threat landscape, generative AI has transformed social engineering from a manual craft into an industrialized process. Cybercriminals can now execute phishing, vishing (voice phishing), and deepfake attacks at a scale and sophistication previously unimaginable.

The stakes are high: Recent data show successful deepfake attacks can result in six-figure losses, with the potential for multi-million dollar fraud. Beyond the balance sheet, these attacks can inflict reputational damage that erodes stakeholder confidence, customer loyalty, and trust. They're also proliferating at a staggering rate. A 2025 Gartner survey found that 62% of organizations had experienced at least one deepfake attack in the last 12 months.

When it's only a matter of time before leaders are misled, the extent of damage will be determined by the strength of their defenses. A multi-layered framework that encompasses cybersecurity, advanced technology, human firewalls, and robust governance is no longer optional.

The industrialization of deception

Traditional phishing attacks were often betrayed by poor grammar and generic greetings, or else caught by spam filters. But today's malicious actors can perfectly mimic voices, faces, and writing styles with AI tools.

AI automates the entire attack lifecycle, from reconnaissance to content creation. It can scrape public data from social media and corporate websites to build detailed target profiles, then leverage that data to craft hyper-personalized messages that reference specific projects, colleagues, or personal events.

Cybercriminals can perpetrate voice impersonation fraud — or vishing — with high-fidelity voice clones of senior executives that are used to call finance employees and issue urgent, seemingly legitimate instructions to execute wire transfers. With just seconds of audio, a voice can be cloned with remarkable accuracy.

Lessons from the trenches of digital impersonation

Analyzing real-world incidents provides actionable lessons for navigating this fraught moment.

In a 2024 masterclass in modern social engineering, a finance worker was tricked into transferring $25 million based on a fraudulent, deepfake-laden video conference. The fraud's success hinged on its layered complexity; attackers didn't just fake one person, they orchestrated an entire video call populated with deepfake versions of the company's CFO and other senior staff. The key takeaway? Standard advice like "verify with a colleague" becomes obsolete when that colleague's digital presence is also a fabrication.

In this case and many others, firms mistakenly trust digital communications and lack backup checks for high-risk transactions. There's no foolproof process that assumes a digital message could be fake.

But case studies also reveal a fundamental weakness in AI-driven attacks: they are constructed from publicly available data. A deepfake can only replicate a public persona.

The most robust defense, therefore, is to create verification steps that rely on private, shared knowledge that an attacker cannot scrape or synthesize. This was demonstrated by a recent attack against the luxury carmaker Ferrari, which foiled attackers' plans because employees relied on knowledge that a deepfake could not possess.

A framework for resilient defense

More than a technical requirement, a holistic defense strategy is now critical to business continuity and financial risk management.

Pillar I: Advanced enterprise identity verification

Advanced threats often succeed by exploiting basic weaknesses, making cybersecurity hygiene the first line of defense. This includes patch management to prevent the initial foothold attackers need to orchestrate an impersonation and multi-factor authentication (MFA), which can reduce risk by over 99%. In the event of a compromise, Zero Trust principles and identity architectures further limit attackers' ability to cause damage.

Pillar II: Advanced countermeasures to fight AI with AI

The second pillar involves embracing advanced technology to fight AI with AI. AI-powered detection tools can identify subtle statistical anomalies in synthetic media that are imperceptible to humans. Through real-time visual, audio, and even physiological analysis, these tools can detect the absence of blood flow or natural human micro-expressions in a video feed — signaling deepfake technology.

Pillar III: Psychological security engineering

Because technology alone is insufficient, as part of the third pillar organizations must harden their human firewall. Leaders should mandate that all high-value transactions or sensitive requests — even those from the CEO — are validated using a pre-agreed, dynamic "secret question" or passphrase that is not recorded in any public or accessible digital system. Organizations can also use AI-powered simulation platforms to conduct realistic training exercises, such as simulated voice impersonation calls with cloned executive voices, to condition employees to pause and verify.

Pillar IV: Deepfake response playbooks

Finally, organizations must turn their attention to governance, developing a deepfake-specific incident response playbook. To ensure a rapid, coordinated, and multi-faceted response, this playbook must include pre-defined workflows for forensic validation, technical containment, crisis communications, and engagement with law enforcement.

Team/Role	Immediate actions (First 60 minutes)	Follow-up actions (1-24 hours)
Security Operations	Isolate systems. Block malicious IPs. Initiate forensic analysis.	Conduct detailed log analysis. Hunt for related IOCs.
IT/Infrastructure	Reset compromised credentials. Verify system integrity.	Accelerate patching of any related vulnerabilities.
Legal & Compliance	Engage counsel. Preserve evidence.	Assess regulatory notifications. Issue content takedown requests.
Corporate Comms/PR	Activate crisis plan. Issue holding statement. Monitor social media.	Draft and issue detailed public statements. Manage media.
Finance Department	Contact banks to halt/recall fraudulent transfers.	Conduct a full audit. Strengthen financial transfer protocols.

Winning the AI arms race

The emergence of generative AI marks the beginning of a perpetual technological arms race between cybercriminals and global enterprises. To withstand the threat of AI-powered deception, leaders must move beyond a reactive security posture and actively verify authenticity. The future of corporate security will be defined not by the ability to block every attack, but by the capacity to maintain operational integrity in a world where reality itself has become the primary attack surface.

This new threat demands more than better tools — it requires a fundamental reset of how organizations authenticate communications, govern risk, and train people to respond. Four strategic priorities now define enterprise resilience in the age of AI-powered deception:

Embrace a zero trust architecture: Shift from a perimeter-based model to one that assumes no user or communication is inherently trustworthy. Continuously and rigorously verify every access request and sensitive transaction.

Invest in a layered defense: There is no silver bullet. A robust defense integrates network security, endpoint protection, content verification, and strong authentication protocols so that a failure in one layer is caught by another.

Cultivate a culture of healthy skepticism: The most resilient organizations will be those where every employee is conditioned to pause and verify unusual or urgent requests, regardless of their apparent source.

Limit your public data footprint: The raw material for deepfakes is public data. Advise executives and key personnel to be mindful of the high-quality audio and video content they share online. Limiting the training data makes it harder for attackers to create convincing fakes.

kyndryl