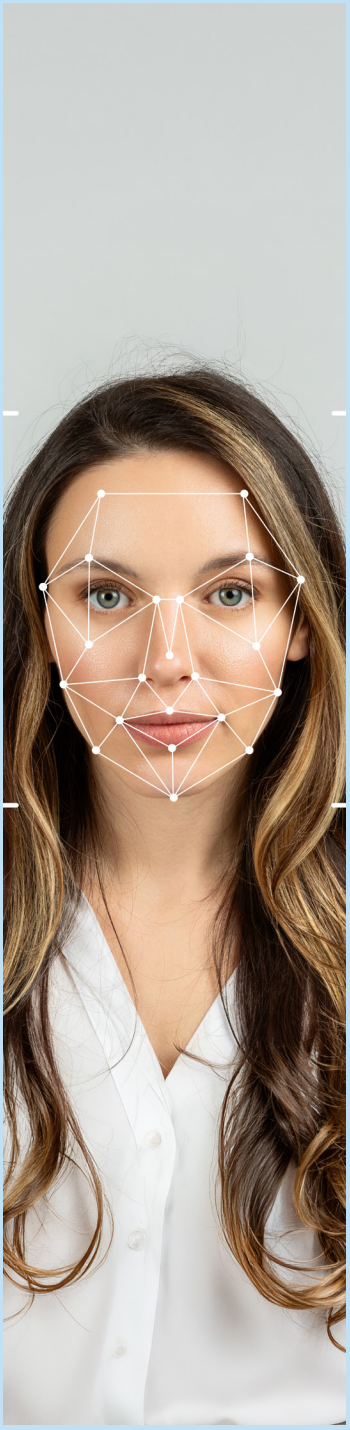
VIEW POINT

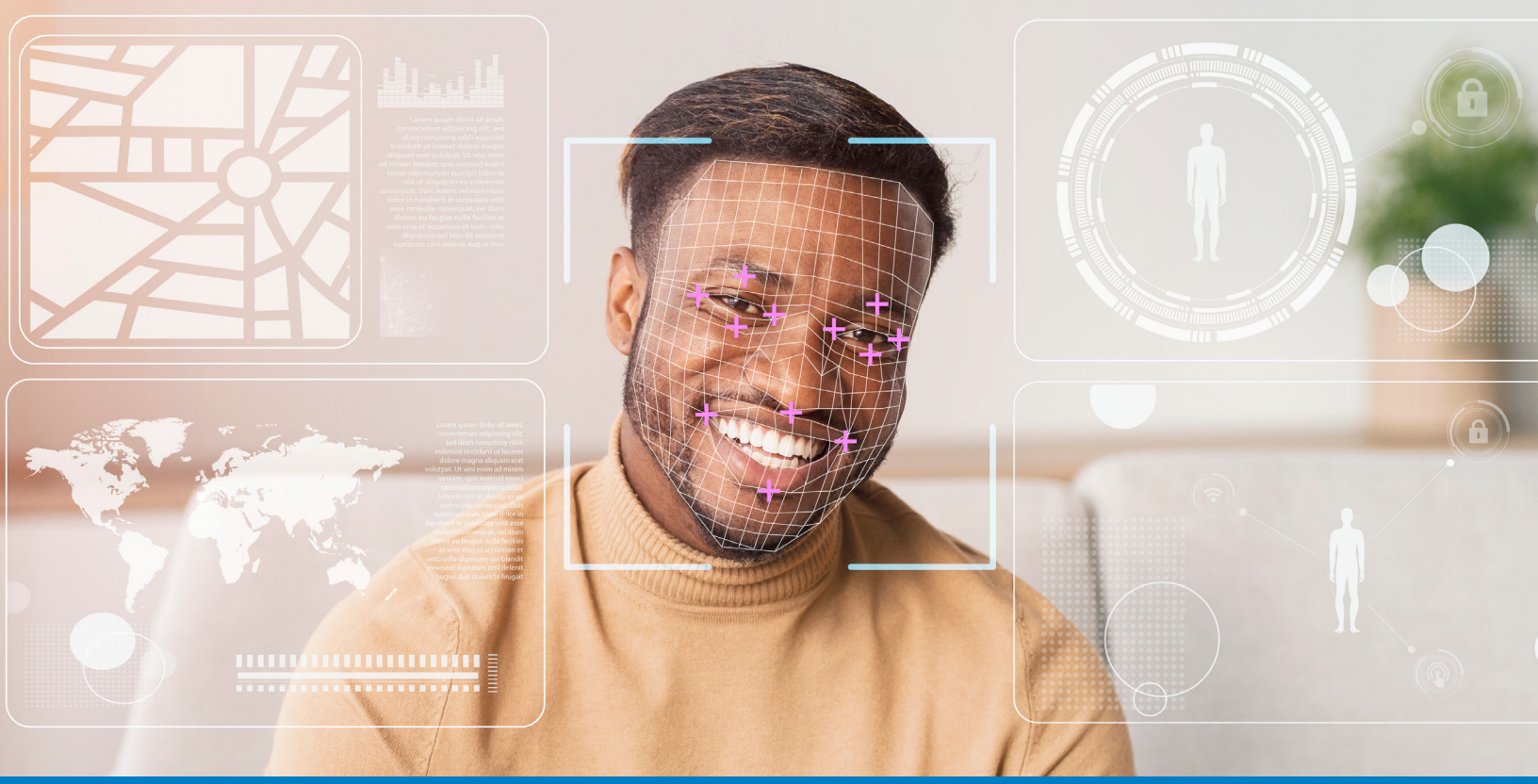


COMBATING DEEPFAKES IN DIGITAL PLATFORMS: STRATEGIES FOR 2026 DETECTION AND MITIGATION

Abstract

As synthetic media becomes more sophisticated and scalable, deepfakes are emerging as a systemic threat to digital trust, enterprise security, and public discourse. This article examines forward-looking strategies for combating deepfakes in 2026, focusing on advanced deepfake detection techniques, the expanding role of presentation attack detection, and the evolution toward multimodal intelligence frameworks. It outlines how organisations can operationalise real-time monitoring, hybrid human-AI oversight, and governance-aligned response mechanisms to mitigate risk. By integrating adaptive detection architectures with ethical safeguards and regulatory alignment, enterprises and platforms can build resilient defence systems capable of preserving integrity in an increasingly AI-driven digital ecosystem.

Infosys®
Navigate your next



As the first quarter of 2026 ends, we see evidence that soon deepfakes, hyper-realistic synthetic media generated using advanced AI models, will graduate from being a fringe threat to an overwhelming cybersecurity concern. As generative systems become more accessible, scalable, and sophisticated, synthetic audio, video, and imagery are reshaping the risk landscape for digital platforms, enterprises, and public institutions alike.

Once associated primarily with manipulated celebrity videos, deepfakes now underpin highly targeted fraud schemes, misinformation campaigns, identity impersonation attacks, and executive spoofing. Deepfake-enabled fraud has already generated billions in global losses, with projections expected to reach USD 40 billion by 2027 as synthetic media becomes more widespread and sophisticated. This escalation is intensifying concerns about vulnerabilities in biometric systems and the resilience of digital identity infrastructure.

This article presents forward-looking, actionable strategies for enterprises, platforms, regulators, and technologists to detect, mitigate, and govern deepfake threats in 2026. It covers the latest deepfake detection techniques, explores the role of Presentation Attack Detection (PAD) in combating synthetic media, and highlights why cross-modal analysis will be essential for future defence systems.

Understanding the scale of this threat is the first step toward building resilient defence strategies.

The deepfake escalation: Why 2026 is a tipping point

Reports indicate that content moderation systems are under growing strain as synthetic media production expands at scale. What was once sporadic manipulation is increasingly automated and industrialised.

The implications are substantial.

- **Misinformation and political manipulation:** Fabricated media alter public perception and undermine democratic discourse.
- **Identity erosion:** Executive impersonation weakens trust even in authenticated environments.
- **Fraud and impersonation:** AI-generated voice or video manipulates employees into financial transfers or data disclosure.

In a high-profile corporate fraud case, criminals used a deepfake impersonation of a company's CFO during a video call to convince a finance employee to authorise a USD 25 million transfer. The incident illustrates how synthetic media can bypass traditional verification controls and exploit executive trust.

To respond effectively, detection capabilities must evolve beyond isolated artefact spotting toward systems capable of multi-layer signal analysis and adaptive model refinement.

The technical foundation: Modern deepfake detection architectures

Deepfake detection techniques form the technical foundation of defence. Over recent years, methods have evolved from simple statistical inconsistencies to sophisticated AI-driven systems capable of analysing visual, audio, and temporal artefacts simultaneously. Core detection models

As generative models grow more advanced, detection systems increasingly rely on Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), Multimodal Large Language Models (MLLMs), Recurrent Neural Networks (RNNs), and other transformer-based models.

These architectures analyse:

- Subtle pixel-level inconsistencies
- Facial micro-movements
- Temporal coherence across frames
- Audio-visual synchronisation

By extracting complex spatial and temporal signals, these models improve classification accuracy across increasingly sophisticated synthetic media environments.

Explainability as a detection requirement

Accuracy alone is no longer sufficient. Detection systems must also be explainable. Future frameworks combine deep learning models with interpretable techniques such as feature attribution and contribution scoring. These tools allow stakeholders to understand why content has been flagged.

Explainability strengthens institutional trust, supports accountable oversight, and provides defensible justification in legal or forensic contexts.

From single-channel analysis to cross-modal intelligence

Detection is evolving from isolated signal analysis to cross-modal intelligence. Next-generation architectures use cross-attention mechanisms and dynamic embeddings to unify audio, visual, and contextual signals. By synchronising insights across modalities, these systems move beyond single-channel detection toward holistic integrity assessment.

Extending identity security: Presentation Attack Detection (PAD) in the synthetic era

Lessons from biometric security, particularly PAD, offer valuable insight for combating synthetic impersonation.

PAD was originally designed to detect biometric spoofing, such as printed photos, masks, or replayed recordings. By verifying liveness, behavioural cues, and sensor-level authenticity at the point of capture, PAD distinguishes genuine human presence from manipulated inputs.

As deepfake techniques become more convincing, PAD capabilities are increasingly being explored and adapted to counter real-time impersonation and synthetic identity attacks.

Two principles are central:

- **Dynamics over static artefacts:** Modern PAD systems analyse micro-movements, subtle reflections, and natural motion coherence, signals generative systems struggle to replicate consistently at scale.
- **Multi-sensor validation:** By combining texture analysis, motion tracking, and depth sensing, PAD strengthens resilience against spoofed presentations.

PAD reinforces trust at the presentation layer. However, as synthetic threats expand beyond spoofed capture to fully fabricated media compositions, defence must extend beyond the point of interaction. While they significantly strengthen deepfake detection techniques, their real-world impact depends on how effectively they are operationalised within platforms and enterprise environments.

Operationalising detection: Embedding deepfake defence into enterprise systems

Platform operators, enterprises, and digital service providers must embed deepfake defence into broader cybersecurity, fraud management, compliance, and governance frameworks.

Continuous monitoring and escalation pipelines

Synthetic media requires continuous monitoring rather than periodic review. Organisations should integrate detection capabilities into:

- Content moderation systems
- Identity verification workflows
- Fraud detection platforms
- Internal communication channels

When suspicious activity is detected, systems should trigger predefined escalation pathways, such as content flagging, secondary verification checks, or alerts to security teams. Rapid detection must be matched by rapid containment to reduce reputational, financial, and operational impact.

Human and AI collaboration

Automation provides scale, but emerging threats demand contextual judgment. Effective defence models combine AI-driven screening with structured human oversight.

This layered approach:

- Reduces false positives
- Identifies novel manipulation patterns
- Continuously improves model performance through feedback loops

By blending machine efficiency with human discernment, organisations strengthen resilience while maintaining fairness and user trust.

Structured incident response

Deepfake incidents should be incorporated into formal incident response plans. Organisations must define:

- Clear cross-functional ownership
- Notification protocols
- Documentation requirements
- Regulatory alignment procedures

Prepared response frameworks enable coordinated action rather than reactive crisis management. Governance integration is essential to limit financial loss and manage evolving legal exposure.

Governance and standards: Aligning policy with technical capability

As enterprises operationalise detection systems, regulatory ecosystems are evolving in parallel. The United Nations' International Telecommunication Union (ITU) has called for stronger global measures to address multimedia authenticity, including verification tools, provenance mechanisms, and standardised labelling approaches for AI-generated media.

Legislative developments, such as the United States' TAKE IT DOWN Act, demonstrate growing expectations for platform accountability in cases involving non-consensual synthetic content.

Emerging evaluation frameworks are expected to guide organisations in assessing deepfake detection systems against evolving security and authenticity principles. Policymakers and standards bodies face the challenge of balancing innovation, freedom of expression, and protection against deception.

Responsible deployment: Ethics, transparency, and accountability

As deepfake detection systems become embedded in digital infrastructure, defence strategies must prioritise transparency and interpretability.

Detection mechanisms should provide understandable explanations for flagged content. Clear review and appeal pathways help ensure contested decisions can be reassessed through structured oversight.

Explainability is central to responsible deployment. Transparent decision logic reduces bias risk, protects legitimate creators from wrongful classification, and supports forensic analysis where legal implications arise.

Strengthening human resilience: User awareness and trust calibration

After this watershed moment in synthetic media, deepfake defence will extend beyond technical safeguards and governance controls to encompass a critical human dimension. Social engineering remains a dominant breach vector: over 98% of cyber incidents involve human factors such as phishing or impersonation. With deepfakes capable of mimicking trusted voices and faces in real time, this vulnerability will become even more consequential.

Forward-looking organisations will embed deepfake readiness into enterprise training programmes, moving from one-off awareness efforts to continuous, scenario-based simulations. Tailored exercises for finance, HR, legal, and executive teams will help employees recognise emerging manipulation tactics before they materialise into fraud or data loss. Identity deception and business email compromise already drive billions in annual fraud losses globally.

Equally important will be proactive trust calibration. As synthetic media becomes more convincing, platforms must provide clearer labelling, contextual signals, and transparent explanations of detection outcomes to help users interpret content responsibly. In an era of accelerating synthetic realism, resilience will depend on cultivating informed, vigilant human participants alongside advancing technical defences.

The co-evolution frontier: Detection in an adversarial AI ecosystem

Deepfakes and detection technologies exist in a continuous evolutionary cycle. As generative systems improve, defensive capabilities must adapt just as rapidly.

Future defence strategies are likely to focus on:

- **Adversarial training:** Strengthening models by exposing them to cutting-edge synthetic content during development.
- **Hybrid analytical systems:** Combining neural network outputs with cross-modal reasoning techniques.
- **Federated and privacy-preserving approaches:** Advancing distributed learning models that enable collaborative improvement without centralising sensitive data.

These approaches reflect a broader shift: detection systems must continuously co-evolve alongside generative technologies.

Strategic priorities for 2026 and beyond

Resilience in 2026 will depend on embedding deepfake mitigation into core digital infrastructure rather than treating it as a standalone control. Key priorities include:

- **Continuous, risk-based monitoring:** Correlating behavioural, biometric, and contextual signals in real time to detect synthetic manipulation early.
- **Integrated identity and fraud controls:** Embedding deepfake detection within onboarding, authentication, and transaction verification processes.
- **Transparent and explainable detection systems:** Providing clear reasoning behind automated decisions to support accountability and user trust.
- **Harmonised evaluation standards:** Developing shared benchmarks and authenticity frameworks to enable interoperability and consistent assurance.
- **Operational preparedness:** Incorporating synthetic threat scenarios into incident response exercises, executive oversight, and enterprise risk modelling.
- **Cross-sector collaboration:** Strengthening cooperation among technology providers, enterprises, policymakers, and standards bodies to address evolving threats collectively.

Sustained investment in adaptive, explainable, and collaborative defence mechanisms will define digital trust in the synthetic era.

deepfake defence as a strategic imperative

In 2026 and beyond, deepfake threats will be pervasive and persistent. They will influence identity fraud, enterprise risk, political discourse, and societal trust.

Combating them requires a systemic approach that integrates:

- Technical innovation
- Identity-layer safeguards
- Operational integration
- Governance alignment
- Ethical transparency

Deepfake detection techniques, particularly when strengthened by presentation-layer security and cross-modal analysis, are essential building blocks. Effective mitigation depends on real-time capability, structured governance, and collaborative global action.

Enterprises and platforms that act decisively will not only protect their operations and users, but they will also help shape a digital ecosystem where integrity and trust endure despite the rise of synthetic media.

Trust will belong to those who invest in defending it.

For more information, contact infosysbpm@infosys.com

Infosys®
Navigate your next