International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

Deepfakes and Generative AI: Challenges for Digital Trust and Information Security

Apoorva Belur
3rd Year BCA, BMS College of Commerce and Management, Bengaluru, India Email:belurapoorva8@gmail.com

Abstract - Generative artificial intelligence has made it possible to create highly realistic synthetic images, videos, and audio. Although these capabilities support creativity, accessibility, education, and entertainment, they can also be misused for identity impersonation, fraud, harassment, political manipulation, and the circulation of misleading information. Such misuse challenges digital trust: the confidence people place in online identities, communications, and media. This paper examines the relationship between deepfakes, generative AI, digital trust, and information security through a focused review of recent research and relevant technical standards. It discusses common forms of deepfake manipulation, the risks they create, and the strengths and limitations of detection approaches. The review finds that no single technical control can reliably establi sh authenticity in every context. Detection systems may struggle with unfamiliar generation methods, compression, and changing content formats; provenance systems can help establish the origin and editing history of media but do not independently prove that every claim in the media is true. A layered approach is therefore needed, combining detection, content provenance, secure identity verification, platform response, user education, privacy safeguards, and proportionate legal and organisational measures. The paper proposes a practical framework for improving digital trust while recognising that the effectiveness of individual measures must be evaluated in real-world settings. As this is a literature-based paper, it does not report original experiments or claim measured detection performance.

Keywords: Deepfakes, Generative AI, Digital Trust, Information Security, Cybersecurity, Misinformation, AI Detection

1. INTRODUCTION

Digital communication increasingly depends on images, videos, voice messages, online profiles, and recorded meetings. People often treat these media as evidence of identity or events. Generative artificial intelligence (AI), particularly deep- learning-based image, audio, and video generation, can produce or modify such media in ways that are difficult for an ordinary viewer to recognise. The resulting content is commonly described as a deepfake when it realistically imitates or manipulates a person's appearance, voice, or actions.

Deepfake technology is not inherently harmful. It can support film production, dubbing, accessibility, creative expression, training, and educational demonstrations. The risk arises when synthetic media is presented without context or used to deceive, impersonate, intimidate, or manipulate. A convincing voice clone may be used in a fraudulent request for money; a fabricated video may be circulated as evidence of an event that never occurred; and non-consensual synthetic intimate imagery can cause serious personal harm.

Digital trust is broader than whether one image is real or fake. It includes confidence in the identity of a speaker, the integrity and origin of a media file, the reliability of a platform, and the processes used to verify information. Deepfakes place pressure on each of these elements. This paper reviews the main risks and mitigation approaches and argues that information security must combine technical safeguards with human judgement, transparent governance, and responsible communication.

2. PROBLEM STATEMENT

The central problem is that realistic synthetic media can weaken confidence in digital evidence and enable security, privacy, and information-integrity harms. Users may be unable to distinguish authentic media from manipulated content, while automated detectors may not generalise to every new generator, format, or distribution environment. At the same time, treating all unfamiliar or disputed media as fake can also damage trust and silence legitimate evidence.

The challenge is therefore not simply to build a detector that labels a file as real or fake. It is to develop a dependable, privacy- conscious process for assessing media authenticity, verifying identity, communicating uncertainty, and responding proportionately when manipulation is suspected.

3. OBJECTIVES

- To explain deepfakes and their relationship with generative AI.
- To identify major risks to digital trust, privacy, cybersecurity, and information integrity.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

• To review key families of deepfake detection and content-authenticity techniques.

• To examine the limitations of current technical and human-centred approaches.

• To propose a layered framework for responsible mitigation and future research.

4. SCOPE AND CONTRIBUTION

This paper focuses on synthetic and manipulated images, videos, and audio, with particular attention to their implications for digital trust and information security. It considers detection models, identity verification, provenance standards, platform processes, media literacy, and organisational response. The discussion is based on published review literature and technical documentation rather than a new dataset or experimental system.

The contribution is a practical synthesis: it distinguishes detection from provenance, authenticity from truthfulness, and technical controls from governance measures. These distinctions are important because a file may have a verifiable editing history yet contain a misleading claim, while an authentic recording may be taken out of context.

5. RELATED WORK AND BACKGROUND

5.1. Deepfakes and Generative AI

Deepfakes can involve face swapping, facial reenactment, synthetic image generation, voice cloning, lip synchronisation, or combinations of audio and video manipulation. Modern generative systems can reduce visible artefacts that older detectors relied on. Abbas and Taeihagh (2024) review deepfake generation and detection approaches and identify technical and policy challenges associated with mitigating disinformation.

5.2. Deepfake Detection Research

Detection research uses visual artefacts, facial and biological cues, temporal inconsistencies, audio features, frequency- domain signals, and multimodal relationships between sound and video. Kaur et al. (2024) identify recurring issues such as limited or imbalanced training data, computational cost, reliability, and generalisation to unseen manipulations. Heidari et al. (2024) likewise discuss limitations around generalisation and interpretability. These findings suggest that performance reported on a benchmark should not automatically be assumed to transfer to every real-world platform or newly generated sample.

5.3. Provenance and Content Credentials

Content provenance records information about where a digital asset came from and how it has been created or modified. The Coalition for Content Provenance and Authenticity (C2PA) specifies a way to associate media with signed provenance information, allowing compatible tools to validate whether the associated claims have been altered (C2PA, 2025). Provenance can provide useful context, but it is not a universal truth detector: it cannot independently guarantee that the depicted event happened as described, and the absence of credentials does not by itself prove manipulation.

5.4. Digital Trust and Information Security

Information security traditionally protects confidentiality, integrity, and availability. Deepfakes add a human-facing dimension to integrity: people must judge whether a message, identity, or recording is genuine. Security controls such as multifactor authentication, callback verification, access controls, incident reporting, and audit trails remain relevant because a synthetic voice or video should not be treated as sufficient proof of identity for a high-risk transaction.

6. RESEARCH METHODOLOGY

This paper uses a focused narrative literature-review approach. Sources were selected for relevance to deepfake generation and detection, the reliability of detection systems, and content provenance. The source base includes peer-reviewed review articles published in 2024 and the C2PA technical specification. The materials were compared thematically across four areas:
(1) types of manipulation and associated harms, (2) detection techniques and limitations, (3) provenance and identity verification, and (4) social, organisational, and policy responses.

The analysis is qualitative and conceptual. No participants were recruited, no original dataset was collected, and no detector was trained or benchmarked. Consequently, the paper does not report original accuracy scores or claim that a particular mitigation method has been experimentally validated by this study. A future empirical extension could compare detector performance across different generators, platforms, compression levels, and audio-visual conditions.

7. MAJOR THREATS TO DIGITAL TRUST

7.1. Misinformation and Public Confusion

Fabricated or altered media can be circulated as evidence of statements, incidents, or public events. Even when a false clip is later

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

corrected, its initial distribution may shape perceptions. Repeated exposure can also produce general suspicion toward genuine recordings, sometimes described as the "liar's dividend": a person can dismiss authentic evidence by claiming it is synthetic.

7.2. Identity Impersonation and Financial Fraud

Cloned voices and synthetic video can be used to impersonate family members, employees, executives, or service representatives. A manipulated call may pressure a recipient to transfer money, disclose credentials, or bypass an established approval process. Organisations should therefore verify high-risk requests through a separate trusted channel rather than relying on voice or video alone.

7.3. Privacy, Harassment, and Consent

A person's face or voice may be reproduced without permission, including in humiliating or intimate content. The consequences can include reputational damage, emotional distress, threats, and unwanted exposure. Effective responses must prioritise the affected person's safety, privacy, and control over further distribution.

7.4. Cybersecurity and Social Engineering

Deepfakes can make phishing and social-engineering attempts more persuasive by imitating a trusted individual. The media itself may not compromise a computer system, but it can manipulate a human decision that grants access, reveals information, or authorises a transaction. Security awareness and strong operational controls remain essential.

7.5. Erosion of Confidence in Digital Evidence

When people believe that any recording can be fabricated, they may distrust legitimate journalism, evidence, and communication. This creates a dual risk: accepting false media and rejecting genuine media. A responsible response must avoid overconfident accusations and should communicate what is known, what remains uncertain, and how the conclusion was reached.

8. DETECTION AND MITIGATION TECHNIQUES

8.1. AI-Based Detection

Machine-learning systems can classify media using visual artefacts, facial movements, inconsistencies between audio and video, spectral patterns, or other learned features. Multimodal approaches can compare whether a speaker's lip movements align with the audio and whether signals from different modalities are consistent. Detection tools can support triage, but their outputs should be treated as evidence with uncertainty rather than unquestionable verdicts.

8.2. Human and Contextual Verification

Reviewers can examine the original source, publication history, surrounding context, corroborating reports, metadata, and whether the same event is documented by independent sources. Reverse-image search and frame extraction can help trace earlier appearances of visual material. Human review is not infallible, especially when content is emotionally charged or highly realistic, so it should complement rather than replace technical checks.

8.3. Content Provenance and Watermarking

Signed provenance metadata can indicate the source and editing history of an asset. Watermarks may help identify certain synthetic outputs, but they can be removed, degraded, or absent when content passes through unsupported tools. Provenance, watermarking, and detection serve different purposes and should be combined where appropriate. A valid provenance record supports claims about origin and handling; it does not certify that the content is fair, complete, or factually true.

8.4. Strong Identity and Transaction Controls

For sensitive actions, organisations should require controls independent of media appearance or voice. Examples include multifactor authentication, verified callbacks to known numbers, dual approval for payments, transaction limits, and documented escalation procedures. Employees should be trained to recognise urgency, secrecy, and requests to bypass normal process as warning signs.

8.5. Platform and Incident Response

Platforms can provide accessible reporting channels, preserve relevant evidence where lawful, label synthetic media when reliable information is available, and reduce harmful distribution under clear policies. Responses should account for satire, consent-based creative work, journalism, and legitimate expression. False positives and mistaken takedowns can themselves harm trust, so decisions need review and appeal mechanisms.

9. PROPOSED LAYERED FRAMEWORK FOR DIGITAL TRUST

This paper proposes a layered framework rather than dependence on a single detector. The layers are complementary and should be selected according to the risk and context of the content.

- Layer 1 — Prevention: establish consent rules, secure accounts, minimise unnecessary access to biometric material, and usesecure workflows for sensitive communications.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

- Layer 2 — Verification: confirm identity through independent channels; verify unusual requests using established contactdetails and approval procedures.
- Layer 3 — Content assessment: combine automated detection, source and context checks, metadata review, and humananalysis. Record uncertainty and avoid binary conclusions when evidence is insufficient.
- Layer 4 — Provenance: adopt interoperable signed provenance and watermarking where practical, while explaining thatmissing credentials are not proof of fakery.
- Layer 5 — Response and accountability: provide reporting and appeal processes, preserve evidence responsibly, supportaffected people, and document decisions.
- Layer 6 — Education and evaluation: teach users how to verify claims, update procedures as generation methods change, andassess controls against new and previously unseen examples.

Figure 1. Layered approach to digital trust: Prevention → Verification → Content Assessment → Provenance → Response and Accountability → Education and Evaluation. The layers form a continuous cycle; evaluation feeds improvements back into prevention and verification.

The framework is designed to be risk-based. A casual entertainment clip may require contextual labelling, while a financial transfer, legal claim, or public-safety announcement requires stronger independent verification. No single layer should be presented as a guarantee of authenticity.

10. ETHICAL, LEGAL, AND GOVERNANCE CONSIDERATIONS

Mitigation must balance safety with privacy, fairness, and freedom of expression. Collecting large amounts of face or voice data to train detectors may introduce privacy and consent risks. Detection systems can also perform unevenly across demographic groups, recording conditions, languages, and devices. Evaluations should therefore examine false positives and false negatives across relevant settings and avoid presenting a model's output as definitive proof of misconduct.

Organisations should define responsibility for labelling, escalation, takedown, evidence retention, and user appeals. Synthetic media created with consent for accessibility, entertainment, or education should not automatically be treated like malicious impersonation. Conversely, the use of a tool for creative purposes does not remove the need to address consent, disclosure, and foreseeable harm.

Legal and regulatory requirements differ by jurisdiction and evolve as technology changes. Institutions should consult current applicable laws and official guidance rather than relying on a generic technical checklist. Good governance should be transparent, proportionate, regularly reviewed, and attentive to the rights of people whose likeness or voice has been used.

11. DISCUSSION

The literature indicates that deepfake mitigation is an evolving problem involving technical, human, and institutional factors. Detection models are useful, but their reliability can vary when tested outside the datasets and conditions on which they were developed. Provenance can add valuable information about an asset's history, but it cannot answer every question about truth or context. Identity verification and organisational procedures address a different part of the problem by reducing the chance that a convincing impersonation will directly trigger a harmful action.

These findings support a shift from asking only "Is this media fake?" to asking a sequence of questions: Where did it come from? Is the source identifiable? Has the file or its provenance been altered? Is the claim corroborated independently? What is the consequence of acting on it? What level of verification is appropriate? This approach recognises uncertainty without treating uncertainty as proof of deception.

An important research priority is evaluation in realistic settings. Systems should be tested on new generation methods, compressed and re-shared media, varied languages and accents, and cross-platform conditions. Reporting should include more than a single accuracy figure: precision, recall, false-positive rates, calibration, processing time, and performance on unseen sources can all matter depending on the application.

12. LIMITATIONS

This paper is a focused narrative review, not a systematic review with an exhaustive database search or formal quality-scoring protocol. It does not present original experiments, a new detector, a user survey, or measured outcomes from an implemented intervention. The reviewed field changes rapidly, so specific techniques and standards may evolve. The proposed layered framework is a conceptual synthesis and should be validated through field studies before being treated as a proven operational model.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 11, CONFLUENCE - 2026

13. FUTURE SCOPE

Future work can develop standardised, cross-platform benchmarks that include new generation models, compressed media, audio-only impersonation, multilingual samples, and multimodal content. Research should measure not only detection accuracy but also robustness, explainability, computational cost, and performance disparities across relevant groups.

Further work can also study how users interpret labels, provenance indicators, and detector confidence scores. Usable explanations may help people make better decisions, while poorly designed warnings may create false certainty or fatigue. Interoperable provenance systems, privacy-preserving detection, secure identity verification, and coordinated incident- response processes deserve continued evaluation. Educational programmes should focus on practical verification habits rather than expecting the public to identify every deepfake by appearance alone.

14. CONCLUSION

Deepfakes and generative AI create significant challenges for digital trust and information security because they can imitate people, alter perceived evidence, and make social-engineering attacks more persuasive. The risks extend beyond misinformation to identity impersonation, financial fraud, privacy violations, harassment, and declining confidence in genuine digital content.

The review shows that deepfake detection, provenance standards, watermarking, secure identity verification, human review, platform response, and media literacy each address different parts of the problem. None is sufficient on its own. Detection tools can fail on unfamiliar content; provenance indicates origin and editing history but does not prove factual truth; and human judgement can be influenced by context and emotion. A layered, risk-based approach is therefore more appropriate than reliance on a single authenticity score.

Ultimately, digital trust should be supported by transparent evidence, careful communication of uncertainty, privacy - conscious design, and accountable processes. Continued research and practical evaluation are necessary to ensure that safeguards remain effective as generative AI evolves.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)