

Artificial Intelligence and Deepfakes: Visual Manipulation, Digital Trust, and Governance Challenges in Digital Media

Amjed Altabbal

University of Tripoli, (Libya), e-mail: amjedattabbal@yahoo.com



ORCID: <https://orcid.org/0009-0009-5418-4865>

Wassim Korbi

University of Carthage, (Tunisia), e-mail: wassimkorbi@gmail.com



ORCID: <https://orcid.org/0009-0001-2241-1818>

Received: 03/07/2026; Accepted: 11/09/2026, Published: 30/09/2026

Abstract

Artificial intelligence has changed how audio and video are created and shared. Synthetic media can now reproduce a real person's face, voice, and mannerisms with a degree of realism that can make authentication difficult for ordinary viewers. This paper examines deepfakes, defined here as AI-generated or AI-manipulated video, images, and audio, and focuses on their implications for digital trust and credibility. It considers technical developments from generative adversarial networks and StyleGAN-based face generation to diffusion models and voice-cloning systems (Goodfellow et al., 2014; Karras et al., 2019; Karras et al., 2020; Dhariwal & Nichol, 2021). These technologies have legitimate creative and communicative uses, but their malicious use can facilitate political manipulation, identity fraud, non-consensual intimate imagery, and social engineering (Chesney & Citron, 2019; Mirsky & Lee, 2020; Mustak et al., 2023). Ofcom reported in 2024 that 43% of UK internet users aged 16 and over believed they had experienced a deepfake in the first half of the year, illustrating growing exposure to synthetic media. Using a structured narrative review supported by qualitative comparative case-study analysis, this paper examines how deepfakes affect trust and credibility across political, corporate, financial, and personal contexts. The analysis compares cases using four common dimensions: the form of manipulation, the mechanism of trust disruption, the type of harm, and the institutional response. The findings indicate that detection remains important but is insufficient as a stand-alone response. Effective protection requires a layered approach combining detection, provenance, disclosure, institutional verification, platform governance, legal accountability, and media literacy.

Keywords: *Artificial intelligence; deepfakes; digital manipulation; digital trust; credibility; misinformation; provenance; regulation.*

Corresponding Author: Amjed Altabbal

<https://doi.org/10.34118/sej.v8i3.4860>

1. Introduction

Deepfakes have become a significant integrity challenge in digital communication. Advances in generative modeling now allow synthetic video and audio to look and sound convincing enough to compete with authentic recordings in everyday online environments (Mirsky & Lee, 2020; Verdoliva, 2020). As a result, the boundary between documentation and fabrication has become more difficult to assess in politics, journalism, entertainment, and corporate communication (Kietzmann et al., 2020).

The problem is not limited to direct deception. Deepfakes can also increase uncertainty about what should be believed, weakening the shared informational conditions needed for credible journalism and informed public debate. Research on synthetic political video shows that deepfakes can affect perceptions of credibility and increase uncertainty even when viewers are not fully persuaded by a particular clip (Vaccari & Chadwick, 2020; Hameleers et al., 2024). This makes digital trust-not only deception-the central analytical concern of the present study.

Public exposure also appears to be increasing faster than ordinary users' ability to authenticate synthetic content. Ofcom's 2024 research found that 43% of UK internet users aged 16 and over believed they had experienced a deepfake during the first half of 2024, while only about one in ten reported confidence in spotting deepfakes. As shown in Figure 1, the difference between exposure and confidence reinforces the need to consider deepfakes as an information-integrity problem rather than only a technical detection problem.



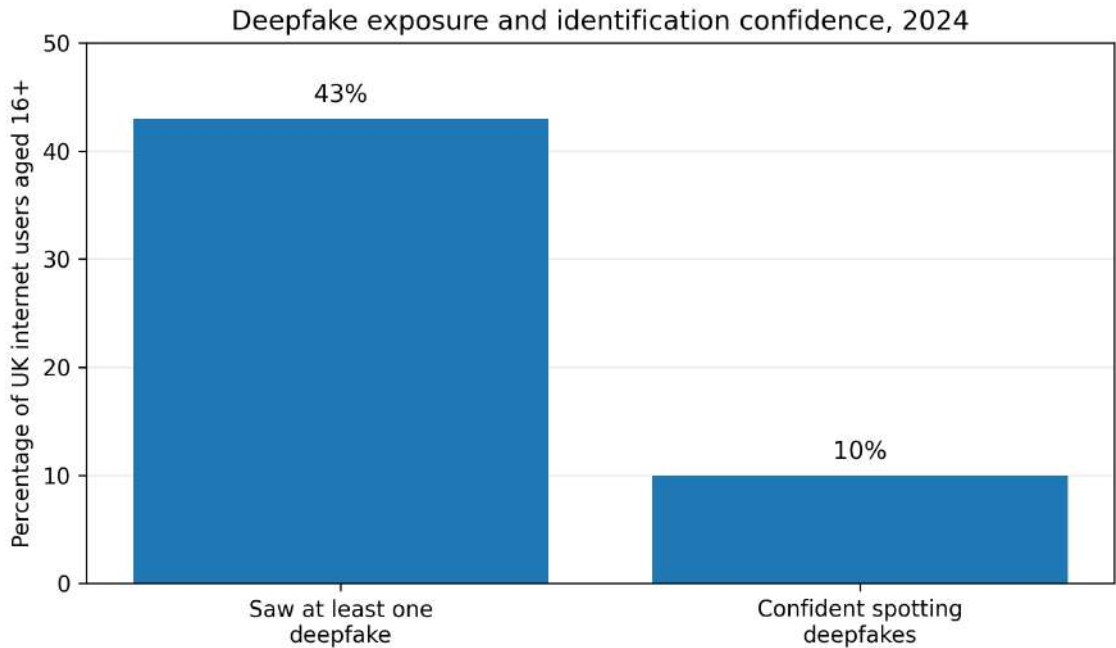


Figure 1. Deepfake exposure and identification confidence among UK internet users aged 16+, 2024. Source: Ofcom (2024).

Recent incidents illustrate the practical consequences. In India, Meta and the Misinformation Combat Alliance launched a WhatsApp helpline and Deepfakes Analysis Unit to help users submit suspicious audio and video for assessment during the 2024 election period (Meta, 2024). In Hong Kong, police described a case in which a finance employee was deceived through a deepfake video conference and authorized transfers totaling HK\$200 million; the incident was subsequently reported as involving Arup (Government of the Hong Kong Special Administrative Region, 2024; KHL Group, 2024). In January 2024, explicit AI-generated images using Taylor Swift's likeness circulated widely online, prompting renewed debate about non-consensual synthetic sexual imagery and platform responsibility (BBC News, 2024; Reuters, 2024).

1.1 Research Problem and Research Gap

Existing research has developed rapidly around deepfake generation, detection, misinformation, privacy, and regulation. However, these strands are often discussed separately. Technical studies emphasize detection performance; communication research examines persuasion, uncertainty, and misinformation; and policy literature focuses on transparency and governance. Less attention has been given to how these strands can be connected through a common analytical focus on digital trust and credibility across different harm contexts. This study addresses that gap by linking the technical problem

Artificial Intelligence and Deepfakes: Visual Manipulation, Digital Trust, and Governance Challenges in Digital Media

of synthetic media with the broader question of how people and institutions decide whether digital evidence is trustworthy.

The contribution is therefore analytical rather than a claim of statistical generalization. The paper uses a structured narrative review to synthesize relevant literature and a comparative case-study approach to examine how the same underlying trust problem appears in different settings. The cases are not treated as isolated descriptions; they are compared using the same analytical dimensions.

1.2 Central Research Question

How do deepfakes disrupt digital trust and credibility across political, corporate, financial, and personal communication contexts, and what combination of technical, institutional, platform, and regulatory responses can reduce these risks?

1.3 Objectives of the Study

- To examine how advances in deepfake generation and detection affect the authenticity and credibility of digital media.
- To analyze how deepfakes disrupt trust across political, corporate, financial, and personal communication contexts.
- To compare selected cases using common dimensions of manipulation, trust disruption, harm, and institutional response.
- To evaluate the strengths and limitations of detection, provenance, disclosure, and organizational verification as countermeasures.
- To examine how emerging regulatory and governance frameworks address the risks associated with synthetic media.

2. METHODS

This study uses a structured narrative review supported by qualitative comparative case-study analysis. The design was selected because the research question requires integration of technical, communication, organizational, and regulatory evidence rather than statistical pooling of effect sizes. The study does not claim to be a systematic review or meta-analysis and does not make statistical generalizations from the selected cases.

The literature search focused primarily on publications from 2018 to 2025, while earlier foundational studies were retained when necessary to explain the development of generative models or established concepts such as the liar's dividend. Searches were conducted through Google Scholar, Scopus, IEEE Xplore, and the ACM Digital Library using combinations of terms such as “deepfake,” “synthetic media,” “GAN,” “diffusion,” “voice cloning,” “detection,” “forensics,” “provenance,” “credibility,”



“trust,” “misinformation,” and “governance.” Official regulatory and institutional documents were also consulted for legal and policy claims, while established news organizations were used only when needed to document specific real-world cases.

Records were screened for direct relevance to at least one of four areas: (1) technical foundations of deepfake generation or detection; (2) effects on misinformation, credibility, trust, privacy, or public communication; (3) economic, financial, or security risks such as impersonation and fraud; and (4) governance, legal, or regulatory responses. Sources were excluded when they were purely promotional, lacked sufficient information to establish the reported claim, or addressed generative AI without a meaningful connection to synthetic media or digital trust.

The case studies were selected purposively to represent distinct but comparable harm contexts: political communication in India’s 2024 election environment; corporate fraud associated with the Hong Kong/Arup video-call incident; non-consensual synthetic sexual imagery involving Taylor Swift; and the broader pattern of AI-enabled CEO voice-impersonation fraud. Case evidence was triangulated using official institutional sources where available and reputable reporting for event documentation.

For the comparative analysis, each case was examined using four common dimensions: (1) form of manipulation; (2) mechanism through which trust or credibility was exploited or disrupted; (3) primary type of harm; and (4) institutional, platform, or regulatory response. This common framework provides a direct link between the literature review, the cases, and the central research question.

Table 1:

Comparative framework for the selected deepfake cases

Case	Domain	Manipulation	Trust/Credibility Mechanism	Primary Harm	Response
India's 2024 election	Politics / disinformation	AI-generated or manipulated audio/video involving political figures and campaign content	Familiarity, political authority, and rapid social-media circulation	Misinformation, voter uncertainty, electoral trust	WhatsApp/MC A helpline and Deepfakes Analysis Unit; verification and fact-checking
Hong Kong / Arup fraud	Corporate finance / cybercrime	Deepfake video conference impersonating executives	Visual and vocal familiarity treated as proof of identity	Financial loss and failure of payment controls	Police investigation; out-of-band verification and multi-person approval
Taylor Swift deepfakes	Privacy / gendered abuse	AI-generated explicit images using a real person's likeness	Visual realism and viral repetition amplified perceived authenticity and reach	Non-consensual sexual abuse, reputational and psychological harm	Platform removal, search restrictions, public and legislative pressure
CEO impersonation scams	Corporate fraud / voice cloning	AI-cloned executive voices used in payment requests	Familiar voice and hierarchical authority encouraged compliance	Fraudulent transfers and social-engineering risk	Warnings from regulators/law enforcement; stronger verification protocols

2.1 Theoretical Framework: Digital Trust, Credibility, and the Evolution of Deepfake Technology

The theoretical framework centers on digital trust and credibility. Deepfakes are technically produced through advances in generative modeling, but their social power depends on the credibility cues people use to judge digital evidence. The framework therefore treats photorealism as an enabling condition rather than the complete explanation of harm. The central analytical proposition is that deepfakes exploit or disrupt trust by making synthetic content appear consistent with familiar people, institutions, and communication contexts.

GANs introduced adversarial training that enabled increasingly realistic synthetic images (Goodfellow et al., 2014). StyleGAN and subsequent refinements improved the realism and controllability of synthetic faces (Karras et al., 2019; Karras et al., 2020). Diffusion models subsequently achieved very high-quality image synthesis, further



narrowing the perceptual distinction between generated and authentic content (Dhariwal & Nichol, 2021). Voice-cloning and neural text-to-speech research similarly demonstrate the ability to reproduce speaker identity from limited reference material (Arik et al., 2017; Jia et al., 2018).

From a trust perspective, the critical issue is that people often use familiar visual and vocal cues as shortcuts for identity and credibility. When those cues can be synthetically reproduced, “seeing” and “hearing” no longer provide sufficient evidence of identity in high-risk contexts. Research on political deepfakes also shows that content does not need to be fully believed to cause harm: uncertainty itself can weaken confidence in information and institutions (Vaccari & Chadwick, 2020; Hameleers et al., 2024).

A related concept is the “liar’s dividend,” whereby the existence of deepfakes can provide an opportunity to dismiss authentic evidence as fabricated. This creates a double risk: false material may be believed, while authentic material may be rejected because audiences know that realistic fabrication is possible (Chesney & Citron, 2019; Schiff et al., 2025).

2.2 Defining Deepfakes: From Faces to Voices

Deepfakes include AI-generated or AI-manipulated images, video, and audio that imitate real people or events. Their forms range from face replacement and reenactment to synthetic speech and multimodal impersonation. Technical surveys emphasize that detection remains difficult because manipulation artifacts may be subtle, model-dependent, and degraded by compression, editing, re-recording, and platform processing (Verdoliva, 2020; Tolosana et al., 2020; Mirsky & Lee, 2020).

For this reason, current research increasingly considers proactive measures in addition to passive detection. Watermarking can provide signals for attribution or authenticity assessment (Zhao et al., 2023), while provenance frameworks such as C2PA seek to preserve information about the origin and editing history of digital content. These approaches do not eliminate the need for detection; rather, they shift part of the integrity problem from identifying manipulation after distribution to documenting content history throughout the media supply chain.

3. RESULTS

The structured narrative review and comparative case analysis point to four connected findings. First, deepfakes create uncertainty that can weaken digital trust even without complete deception. Second, synthetic media can exploit familiar authority cues in financial and corporate settings. Third, detection systems remain vulnerable to



changes in generation, compression, and distribution conditions. Fourth, governance is moving toward layered responses that combine detection with provenance, disclosure, platform responsibility, and institutional controls.

3.1 Societal Impacts and Trust Erosion

Across the reviewed scholarship, a recurring effect is not simply that synthetic media can deceive, but that it makes judgments about authenticity more difficult. Research on synthetic political video shows that deepfakes can influence credibility perceptions and increase uncertainty about news content (Vaccari & Chadwick, 2020; Hameleers et al., 2024). A recent meta-analysis found that human deepfake detection performance was generally close to chance, although training and technological assistance can improve performance (Diel et al., 2024).

The implication is that users may become more skeptical without necessarily becoming more accurate. This distinction matters because excessive skepticism can also damage trust in authentic journalism, evidence, and public communication. The problem is therefore better described as an integrity and credibility challenge than as a simple failure of individual media literacy.

Non-consensual synthetic sexual imagery demonstrates a related form of trust and identity harm. The 2024 circulation of explicit AI-generated images involving Taylor Swift showed how quickly synthetic content can spread and how platform responses can become reactive once material has already achieved substantial visibility (BBC News, 2024; Reuters, 2024).

3.2 Economic and Security Risks

Deepfakes can also strengthen fraud and social engineering because they add realistic audiovisual presence to authority-based scams. In January 2024, Hong Kong authorities described a case in which a finance employee was deceived during a pre-recorded video meeting and authorized transfers totaling HK\$200 million. The incident was subsequently identified as involving Arup, illustrating how familiar faces and voices can be used to exploit organizational assumptions about identity (Government of the Hong Kong Special Administrative Region, 2024; KHL Group, 2024).

Financial regulators have issued parallel warnings. FinCEN has warned that deepfake media can be used in fraud schemes targeting financial institutions and customer verification processes, while the FBI has warned about the use of generative AI to strengthen impersonation and financial fraud. The Bank of Italy has similarly warned about fabricated videos that imitate institutional leaders and attempt to borrow their authority (Financial Crimes Enforcement Network, 2024; Federal Bureau of Investigation, 2024; Banca d'Italia, 2024).



3.3 Detection and Countermeasures

Across the technical literature, detection is a moving target. Detectors may learn artifacts associated with known generation pipelines, but performance can decline when content is compressed, edited, re-recorded, or generated by a new model. This makes generalization under distribution shift a major challenge for deployment outside controlled benchmarks (Verdoliva, 2020; Mirsky & Lee, 2020; Tolosana et al., 2020).

For that reason, countermeasures increasingly combine detection with provenance and disclosure. The C2PA framework provides a mechanism for recording content-origin information, while the EU AI Act establishes transparency obligations for AI-generated or manipulated content, including deepfakes, subject to specified exceptions (Coalition for Content Provenance and Authenticity, 2025; European Union, 2024). Detection should therefore be treated as one component of a broader integrity system rather than as a universal authenticity test.

Human performance evidence also supports a cautious interpretation of automated or unaided judgments. The meta-analysis by Diel et al. (2024) found that overall human detection performance was not significantly above chance, while strategies such as training and AI assistance improved performance. This reinforces the need for human review, institutional verification, and procedural safeguards alongside technical tools.

3.4 Ethical, Legal, and Regulatory Frameworks

The governance evidence shows rapid movement but uneven coverage. UNESCO's Recommendation on the Ethics of Artificial Intelligence emphasizes human rights, transparency, accountability, and human oversight as central principles for responsible AI governance (UNESCO, 2021). In the European Union, the AI Act includes transparency obligations for providers and deployers of certain AI systems. Article 50 requires disclosure when AI-generated or manipulated image, audio, or video content constitutes a deepfake, with specified exceptions for certain artistic, creative, satirical, or fictional contexts (European Union, 2024).

The Digital Services Act adds platform-level obligations concerning systemic risks and due diligence, while the European Commission has supported a structured governance approach to disinformation. These measures indicate a shift from treating deepfakes only as individual pieces of false content toward addressing the systems through which synthetic media is created, labeled, distributed, amplified, and challenged (European Union, 2022; European Commission, 2025).

Artificial Intelligence and Deepfakes: Visual Manipulation, Digital Trust, and Governance Challenges in Digital Media

A cross-cutting result is that transparency and accountability are becoming central governance principles, but enforcement remains difficult because synthetic media can move across jurisdictions and because some harms arise from uncertainty and denial rather than from successful deception alone.

A. 3.5 Comparative Case Studies: Political, Economic, and Personal Harm

1) Case Study 1: India's 2024 Election Deepfakes



Figure 2. Illustrative media coverage of deepfake-related political content in India. Source: image retained from the original manuscript.

India's 2024 election environment highlighted how AI tools can shape political communication and accelerate the circulation of manipulated media. Reporting documented the growing use of AI-generated and manipulated audio and video in political contexts. The significance of the case lies not only in whether a specific clip persuades viewers, but also in the uncertainty created when authentic and synthetic political content become difficult to distinguish.

Responses included verification and reporting workflows rather than reliance on perfect automated detection. Meta and the Misinformation Combat Alliance launched a WhatsApp helpline and Deepfakes Analysis Unit so users could submit suspicious audio and video for assessment (Meta, 2024). This response is consistent with the study's broader finding that digital integrity requires cooperation among platforms, fact-checkers, technical experts, and users.

Analytically, the case demonstrates the interaction between manipulation, trust, harm, and response: synthetic political content exploits recognizable public figures and high-attention communication environments; the resulting harm includes uncertainty and potential misinformation; and the response combines detection, verification, reporting, and public awareness. The case also illustrates the relevance of the liar's dividend because widespread awareness of deepfakes can make it easier to challenge authentic material as well as false material (Chesney & Citron, 2019; Schiff et al., 2025).

2) Case Study 2: The Hong Kong/Arup Corporate Fraud (2024)



Figure 3. Illustrative visual representation of the Arup deepfake corporate-fraud case.
Source: image retained from the original manuscript.

Hong Kong police described a fraud in which a finance employee participated in a pre-recorded video conference that appeared to include the company's chief financial officer and other colleagues. The employee was instructed to execute transactions totaling HK\$200 million. Arup later confirmed that it was the company involved. The case is important because the employee used visual and social cues to assess whether the request was genuine, yet those cues had been synthetically reproduced (Government of the Hong Kong Special Administrative Region, 2024; KHL Group, 2024).

The case demonstrates a direct credibility mechanism: the attackers did not need to defeat every technical security control if they could make the employee believe that legitimate senior executives were present. The practical implication is that high-value financial decisions should not rely on voice or video alone. Out-of-band confirmation, multi-person approval, and established escalation procedures provide stronger safeguards because they introduce an independent verification channel.

The case therefore contributes to the central research question by showing how deepfakes can transform audiovisual familiarity into a vulnerability. The primary harm

was financial, but the deeper institutional consequence was a reduction in confidence that remote audiovisual communication can serve as reliable proof of identity.

The available evidence supports the conclusion that the incident involved convincing synthetic representations of company personnel; however, the precise generation software and technical pipeline have not been established in the sources used here. The analysis therefore avoids attributing the fraud to a specific face-simulation platform without direct evidence.

3) *Case Study 3: Taylor Swift and Non-Consensual Deepfakes (2024)*



Figure 4. Illustrative media representation of the 2024 Taylor Swift deepfake incident.
Source: image retained from the original manuscript.

In January 2024, explicit AI-generated images using Taylor Swift's likeness circulated widely on social media. The incident prompted public criticism, calls for stronger legal protections, and platform action to limit the circulation and searchability of the material (BBC News, 2024; Reuters, 2024).

The case illustrates a different trust mechanism from the political and corporate examples. Here, the central issue is the unauthorized use of a person's identity and the ability of synthetic imagery to create false visual evidence of intimate conduct. The harms include privacy violation, reputational damage, psychological distress, and loss of control over one's likeness. The case also demonstrates the speed mismatch between creation and governance: once synthetic material becomes viral, removal may not fully reverse the harm.

The case therefore extends the paper's trust framework from trust in information to trust in identity. It shows that digital integrity includes not only determining whether a piece of media is authentic, but also protecting individuals from the unauthorized construction and circulation of false representations of themselves.

4) Case Study 4: CEO Impersonation Scams



Figure 5. Illustrative media representation of AI-enabled voice impersonation and executive fraud. Source: image retained from the original manuscript.

Beyond high-profile incidents, industry and law-enforcement reporting documents a wider pattern of AI-enabled CEO impersonation scams. In a widely cited 2019 case, a UK-based energy company executive transferred approximately €220,000 after receiving an AI-generated voice call that appeared to come from the chief executive of the company's German parent organization (Forbes, 2019; MIT Media Lab, 2020).

Unlike the Arup incident, which involved a specific organization and a video conference, the CEO-impersonation pattern demonstrates the scalability of voice cloning. The attack relies on familiarity, hierarchy, urgency, and the assumption that a recognizable voice is a reliable identity signal. This creates an organizational risk because the same social-engineering logic can be repeated across companies and payment environments.

The common lesson across this case and the Arup case is that audiovisual identity should not be treated as sufficient authorization for high-risk transactions. Independent

verification channels and separation of communication from authorization are therefore central controls.

3.6 Cross-Case Comparative Synthesis

The four cases reveal a common mechanism despite their different domains. In each case, synthetic media gains influence by reproducing a credibility cue that people already recognize: political identity, executive authority, personal likeness, or a familiar voice. The harm then depends on the context in which that cue is trusted. Political cases produce uncertainty and misinformation; corporate cases can produce financial loss; personal cases can produce privacy and dignity harms.

The comparison also shows that institutional responses differ in emphasis but share a layered structure. Political cases require rapid verification and public communication; financial cases require independent authorization controls; personal cases require rapid removal and victim protection; and all cases benefit from provenance, disclosure, and public media literacy. This cross-case pattern directly supports the paper's argument that no single detector can address the full integrity problem.

4. DISCUSSION

The results suggest that deepfakes are best understood as a disruption of digital credibility. The technical breakthrough is not only photorealism or voice realism; it is the ability to manufacture social proof. A voice that sounds familiar, a face that appears present, or an image that appears documentary can trigger belief or compliance because audiences routinely use these cues as evidence of identity and authenticity.

The evidence also supports a broader trust dynamic. When deepfakes become common, audiences may respond in two problematic ways: some may accept false content, while others may dismiss authentic evidence because it could be fabricated. This dual effect is central to the liar's dividend and helps explain why deepfakes can damage information environments even when a particular fake fails to persuade everyone (Chesney & Citron, 2019; Schiff et al., 2025).

From a policy and platform perspective, the findings point toward a layered digital-integrity model rather than a detector-only model. The layers are: (1) detection; (2) provenance; (3) disclosure; (4) institutional verification; (5) platform governance; (6) legal accountability; and (7) media literacy. These layers address different points in the lifecycle of synthetic media, from creation and distribution to verification, response, and remediation.

First, disclosure and provenance can make the origin and manipulation history of content more visible. The EU AI Act provides a regulatory basis for disclosure of



deepfake content, while C2PA offers a technical framework for recording provenance when it is adopted across the content ecosystem (European Union, 2024; Coalition for Content Provenance and Authenticity, 2025). Second, platform governance is important because recommendation, reposting, and coordinated amplification can scale harm. Third, institutional security must be redesigned so that high-value decisions do not depend on audiovisual familiarity alone.

Detection remains important, but the reviewed evidence explains why it cannot carry the entire burden. Generators evolve, content is compressed and edited, and different platforms transform media in ways that can alter forensic signals. Human detection also has important limitations. Therefore, the most defensible approach is to treat detection as one instrument within a broader integrity architecture rather than as a definitive authenticity oracle.

Finally, governance must account for unequal harms. Non-consensual intimate deepfakes require rapid takedown pathways, victim support, and legal accountability, while political and corporate contexts require verification procedures that protect democratic communication and financial controls. The cases show that “trust” is not a single outcome: it includes trust in information, trust in identity, trust in institutions, and trust in communication channels.

5. CONCLUSION

The evidence reviewed in this study supports the conclusion that deepfakes are a scalable capability that challenges the assumption that seeing and hearing are sufficient forms of proof. The consequences extend across civic, organizational, financial, and personal life, including political uncertainty, financial fraud, identity impersonation, and non-consensual synthetic imagery.

Detection research has made substantial progress, but real-world media remains difficult to authenticate because content is generated by evolving systems, compressed, edited, reposted, and viewed in contexts that encourage rapid judgment. For this reason, the most credible response is an integrated model that combines detection with provenance, disclosure, platform governance, institutional verification, legal accountability, and media literacy.

The study contributes a comparative trust-centered interpretation of deepfake risk. Across the selected cases, the underlying mechanism is consistent: synthetic media exploits or disrupts credibility cues that people normally use to establish identity and authenticity. Future research should therefore evaluate not only whether detection

systems identify synthetic content, but also whether combined integrity interventions measurably improve trust, reduce harm, and preserve confidence in authentic evidence.

6. REFERENCES

- Arik, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Sengupta, S., & Shoeybi, M. (2017). Deep Voice: Real-time neural text-to-speech. *Proceedings of the 34th International Conference on Machine Learning*, 70, 195–204. <https://proceedings.mlr.press/v70/arik17a.html>
- Banca d'Italia. (2024, May 22). Warning: deepfake videos circulating online. <https://www.bancaditalia.it/media/notizia/warning-deepfake-videos-circulating-online/?dotcache=refresh>
- BBC News. (2024, January 26). Taylor Swift deepfake images spark outrage, <https://www.bbc.com/news/technology-68110476>
- Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6), 1753–1819.
- Coalition for Content Provenance and Authenticity. (2025). C2PA Specifications (version 2.2). <https://spec.c2pa.org/specifications/>
- Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34.
- Diel, A., Lalgı, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. *Computers in Human Behavior Reports*, 16, 100538. <https://doi.org/10.1016/j.chbr.2024.100538>
- European Commission. (2025, February 13). Commission endorses the integration of the voluntary Code of Practice on Disinformation into the Digital Services Act. <https://digital-strategy.ec.europa.eu/en/news/commission-endorses-integration-voluntary-code-practice-disinformation-digital-services-act>
- European Union. (2022). Regulation (EU) 2022/2065 (Digital Services Act). <https://eur-lex.europa.eu/eli/reg/2022/2065/oj>
- European Union. (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act). <https://eur-lex.europa.eu/eli/reg/2024/1689/2026-07-27/eng>
- Federal Bureau of Investigation, Internet Crime Complaint Center. (2024, December 3). Criminals use generative artificial intelligence to facilitate financial fraud. <https://www.ic3.gov/PSA/2024/PSA241203>
- Forbes. (2019, September 3). A voice deepfake was used to scam a CEO out of \$243,000. <https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/>
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S,



- Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.
- Government of the Hong Kong Special Administrative Region. (2024, June 26). LCQ9: Combating frauds involving deepfake. <https://www.info.gov.hk/gia/general/202406/26/P2024062600192.htm>
- Hameleers, M., van der Meer, T. G. L., & Dobber, T. (2024). Distorting the truth versus blatant lies: The effects of different degrees of deception in domestic and foreignpolitical deepfakes. *Computers in Human Behavior*, 152, 108096. <https://doi.org/10.1016/j.chb.2023.108096>
- Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R, López-Moreno, I., & Wu, Y. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in Neural Information Processing Systems*, 31, 4485–4495.
- Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. <https://doi.org/10.1109/CVPR.2019.00453>
- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. <https://doi.org/10.1109/CVPR42600.2020.00813>
- KHL Group. (2024, May 21). Arup employee falls victim to US\$25 million deepfake video call scam. <https://www.khl.com/arup-employee-falls-victim-to-us25-million-deepfake-video-call-scam/8037338.article>
- Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? *Business Horizons*, 63(2), 135–146. <https://doi.org/10.1016/j.bushor.2019.11.006>
- Meta. (2024, February 19). MCA's WhatsApp helpline: Curbing the spread of AI-generated misinformation in India. <https://about.fb.com/news/2024/02/mcas-whatsapp-helpline-curbing-the-spread-of-ai-generated-misinformation/>
- Mirsky, Y., & Lee, W. (2020). The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 54(1), Article 7, 1–41. <https://doi.org/10.1145/3425780>
- MIT Media Lab. (2020, July 21). Deepfakes, explained. <https://www.media.mit.edu/articles/deepfakes-explained/>
- Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., & Dwivedi, Y. K. (2023). Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research*, 154, 113368. <https://doi.org/10.1016/j.jbusres.2022.113368>
- Ofcom. (2024, July 23). A deep dive into deepfakes that demean, defraud and disinform.



<https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/deepfakes-demean-defraud-disinform>

Reuters. (2024, January 29). US lawmakers rally behind Taylor Swift after AI deepfake images spread online. <https://www.reuters.com/technology/us-lawmakers-rally-behind-taylor-swift-after-ai-deepfake-images-spread-online-2024-01-29/>

Schiff, K. J., Schiff, D. S., & Bueno, N. S. (2025). The liar's dividend: Can politicians claim misinformation to evade accountability? *American Political Science Review*, 119(1), 71–90. <https://doi.org/10.1017/S0003055423001454>

Tolosana, R., Vera-Rodríguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64, 131–148. <https://doi.org/10.1016/j.inffus.2020.06.014>

UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. <https://unesdoc.unesco.org/ark:/48223/pf0000380455>

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6(1). <https://doi.org/10.1177/2056305120903408>

Verdoliva, L. (2020). Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5), 910–932. <https://doi.org/10.1109/JSTSP.2020.3002101>

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9(11), 40–53. <https://doi.org/10.22215/timreview/1282>

Zhao, Y., Liu, B., Ding, M., Liu, B., Zhu, T., & Yu, X. (2023). Proactive deepfake defence via identity watermarking. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 4591–4600. <https://doi.org/10.1109/WACV56688.2023.00458>

Citation; Amjed, A. Wassim, K, *Artificial Intelligence and Deepfakes: Visual Manipulation, Digital Trust, and Governance Challenges in Digital Media*. *Social Empowerment Journal*. 2026; 8(3): pp. 22-38. <https://doi.org/10.34118/sej.v8i3.4860>

Publisher's Note: SEJ stays neutral with regard to jurisdictional claims in published maps and institutional affiliations

