Deepfakes Challenging Communication Ecosystems: Governance, Policy, and Detection Guidelines

Anastasiia Iufereva[1], Kasey Miller[2], Filip Beranek[3], Atul Prem Patnala[4] and Karl Ricanek[2]
[1]IULM University, Milan, Italy
[2]University of North Carolina Wilmington, Wilmington, USA
[3]Department of Cybernetics and New Technologies for the Information Society, Plzeň, Czech Republic
[4]Department of Computer Science, University of North Carolina Wilmington, Wilmington, USA

anastasiia.iufereva@studenti.iulm.it
millerkc@uncw.edu
bionothi@ntis.zcu.cz
ap9107@uncw.edu
ricanekk@uncw.edu

Abstract: The rapid advancement of Generative Artificial Intelligence (GenAI) has made deepfakes increasingly realistic, creating significant challenges for communication ecosystems, democratic processes, cybersecurity, and public trust. Existing detection approaches often rely on single-modality techniques that struggle to keep pace with rapidly evolving generative models, while governance and regulatory responses remain fragmented. This study addresses these challenges through the first three phases of the Design Science Research (DSR) methodology: problem explication, requirements definition, and artifact design. Drawing upon a scoping review of communication, cybersecurity, information policy, and artificial intelligence literature, the research identifies key technical and governance requirements for trustworthy deepfake verification. The primary artifact is a multimodal deepfake detection framework that supports both video and still-image analysis by integrating facial analytics, audio-visual consistency validation, metadata and provenance analysis, embedding-based anomaly detection, and an ensemble of specialized image-forensic detectors. The framework combines these complementary sources of evidence to generate calibrated confidence scores and explainable verification results while remaining extensible to emerging generative AI models. Beyond technical detection, the study proposes an integrated governance framework that combines standardized disclosure and provenance mechanisms, institutional verification processes, collaborative governance, and media literacy. The framework is aligned with emerging regulatory initiatives, including the European Union AI Act and the California AI Transparency Act, while identifying persistent gaps in transparency, accountability, and cross-sector coordination. By integrating technical detection with governance and policy, this research contributes a practical framework for strengthening information integrity and public trust while supporting the responsible adoption of generative AI.

Keywords: Deepfakes, Deepfake detection, Generative AI, Design science research, EU AI Act, California AI Transparency Act

1. Introduction

The rapid development of generative artificial intelligence has a significant impact on the media ecosystem. As AI-generated images, videos, and audio recordings become increasingly realistic, it can become much harder for people to distinguish authentic content from fake or manipulated material. This makes synthetic media a growing challenge for the contemporary communication environment (Diel et al., 2025; Hynek et al., 2025; Iufereva, 2024). One of the most controversial forms of synthetic media is deepfakes. Deepfakes are understood as AI-generated or AI-manipulated media content that can realistically imitate people, voices, or events. While deepfake technologies may also have positive outcomes for the education and creativity sphere, they can also create serious issues for political communication, public administration, corporate reputation, and even everyday digital communication (Raemy et al., 2025; Salvi, 2024; Nizami, 2025). For example, recent research shows that deepfakes may undermine confidence in public institutions, intensify disinformation, and contribute to uncertainty about what is real or fabricated (Ahmed et al., 2025; Diel et al., 2025; Hynek et al., 2025).

At the same time, detection technologies are also developing rapidly. Recent studies have moved from single-modality detection toward multimodal approaches. This combines facial analytics, audio-visual consistency, spatial–temporal analysis, and embedding-based anomaly detection (Pei et al., 2026; Prabhu et al., 2026; Raza et al., 2026). These tools may help identify synthetic or manipulated content. However, the detection process remains limited for a few reasons. First, the rapid development of generative AI means that detection methods can quickly become outdated. Second, detection tools often struggle with edited or even slightly modified images, videos, and audio recordings. Third, recent in-the-wild evaluations show that the accuracy of state-of-the-art detectors drops precipitously on contemporary synthetic media, and even leading commercial detection

services fall short of the accuracy of human forensic analysts (Chandra et al., 2026; Li et al., 2025). Therefore, there is a need to develop more effective approaches to AI-generated content verification. It must be embedded within broader verification workflows, governance structures and transparency mechanisms.

Policy responses to deepfakes are also emerging, but they remain uneven. The EU AI Act introduces transparency obligations for AI-generated and manipulated content, including machine-readable marking and disclosure requirements for deepfakes (European Union, 2024). The European Commission's Code of Practice on marking and labelling AI-generated content further develops this focus on transparency and labelling (European Commission, 2025a; European Commission, 2026). In the United States, California's AI Transparency Act introduces disclosure, watermarking and detection obligations for covered providers of generative AI systems (California Legislature, 2024). However, many jurisdictions still rely on fragmented legal measures or general laws related to disinformation, deepfakes and privacy issues.

This study addresses these challenges by examining deepfakes through an interdisciplinary lens. This combines communication studies, information policy, cybersecurity, and artificial intelligence research. The purpose of the study is to develop governance, policy, and detection guidelines to mitigate the risks posed by the proliferation of GenAI in communication ecosystems. The study is guided by the following research questions:

RQ1: How do recent studies conceptualize GenAI (deepfakes) as a challenge to communication ecosystems and public trust?

RQ2: What mitigation approaches are currently discussed, including detection, labeling, provenance, media literacy, platform governance, and regulation?

RQ3: What shortcomings remain in transparency, accountability, and cross-sector coordination?

RQ4: How do regulatory approaches such as the EU AI Act and California AI Transparency Act compare with fragmented governance models?

2. Literature Review

2.1 Deepfakes as a Challenge to Communication Ecosystems

Deepfakes are increasingly understood not only as manipulated content but as a systemic challenge to contemporary communication ecosystems. Advances in generative artificial intelligence have made synthetic image, video, and audio content more realistic, allowing malicious actors to imitate public figures, fabricate events, manipulate political communication, and damage institutional or corporate reputation (Diel et al., 2025; Hynek et al., 2025). As synthetic media circulates through social media platforms, news environments and private messaging systems, it becomes a significant part of broader information flows in which many users are involved (Hastuti et al., 2025; OECD, 2024; UNESCO, 2025a).

The serious issue created by deepfakes is therefore broader than the spread of false content. Deepfakes can weaken the public's ability to evaluate whether digital information is authentic, manipulated, or entirely synthetic (UNESCO, 2025a; Brennan Center for Justice, 2024). This creates risks for democratic and public communication domains, where the credibility of visual and audiovisual evidence is often central to public trust (CIVICUS, 2025; Salvi, 2024; UNESCO, 2025b). For instance, in political contexts, deepfakes may be used to misrepresent candidates, distort electoral information, or amplify disinformation campaigns (CIVICUS, 2025; Ax & Coster, 2026). In organizational and corporate communication, synthetic media may be used to damage reputation, impersonate executives, or generate misleading crisis narratives (Nizami, 2025).

A further challenge is the so-called "liar's dividend," where the existence of deepfakes allows authentic evidence to be dismissed as fabricated (Brennan Center for Justice, 2024). In this sense, deepfakes do not only make false information more convincing; they may also make genuine information easier to deny. This contributes to a broader crisis of trust in digital communication and public institutions (Ahmed et al., 2025; Diel et al., 2025; Hynek et al., 2025). For this reason, deepfakes should be analyzed as a problem of information integrity, verification, and public trust.

2.2 Detection and Verification Approaches

Research on deepfake detection has evolved from identifying visible image artifacts to multimodal approaches that combine visual, audio, temporal, biometric, and contextual information (Pei et al., 2026; Raza et al., 2026; Prabhu et al., 2026). Because contemporary deepfakes often manipulate multiple media modalities simultaneously, integrating diverse evidence generally provides greater robustness than unimodal techniques.

The challenge now extends to both AI-manipulated and fully AI-generated still images. Recent studies show that state-of-the-art detectors perform well on benchmark datasets but experience significant performance degradation under real-world conditions such as compression, resizing, and social-media processing (Cao et al., 2026; Chandra et al., 2026; International AI Safety Report, 2026; Li et al., 2025). Consequently, current research increasingly favors ensembles of complementary detectors, including generative-model fingerprinting techniques, while large vision-language models provide explainable outputs that assist rather than replace human reviewers (Xu et al., 2025; Wen et al., 2025; Yu et al., 2025).

Because generative and detection models continue to evolve in parallel, detection should be viewed as one component of a broader verification process rather than a standalone solution (International AI Safety Report, 2025; Chandra et al., 2024). Hybrid workflows combining automated detection, provenance analysis, and human review provide greater resilience than any single approach, particularly given the risks of false positives and false negatives (Canadian AI Safety Institute, 2025; UK Department for Science, Innovation and Technology, 2026; Romero-Moreno, 2025; Somoray et al., 2025).

2.3 Labelling, Watermarking and Provenance as Transparency Infrastructure

Because deepfake detection remains imperfect, increasing attention has shifted toward transparency mechanisms such as provenance, watermarking, and standardized labeling (Chandra et al., 2024; Coalition for Content Provenance and Authenticity, 2025; European Union, 2024). Provenance records, cryptographic content credentials, and watermarking help establish the origin and modification history of digital content, supporting verification rather than replacing it (Content Credentials, 2025).

Current policy emphasizes both machine-readable and human-readable transparency. Technical mechanisms—including provenance metadata, watermarking, and C2PA content credentials—support automated verification by platforms, journalists, regulators, and fact-checkers (Coalition for Content Provenance and Authenticity, 2025; Content Credentials, 2025; European Union, 2024; European Commission, 2025a). Human-readable labels improve user awareness, although evidence suggests they are most effective when combined with broader verification processes rather than presented as standalone indicators (Al Aghbari et al., 2026; Li & Yang, 2024). Consequently, effective transparency requires integrating technical provenance with clear public disclosure.

2.4 Governance and Policy Responses

Deepfake governance remains uneven across jurisdictions. The European Union has adopted one of the most comprehensive regulatory approaches through the EU AI Act, which requires transparency and disclosure for certain AI-generated content, supported by the European Commission's Code of Practice (European Union, 2024; European Commission, 2025a; European Commission, 2025b; European Commission, 2026). California's AI Transparency Act (SB-942) similarly emphasizes disclosure, watermarking, and detection obligations for generative AI providers (California Legislature, 2024; Future of Privacy Forum, 2024).

Outside these initiatives, governance remains fragmented, with many jurisdictions relying on existing misinformation, privacy, election, or cybercrime legislation rather than dedicated synthetic media policies (OECD, 2024; Romero-Moreno, 2025; National Conference of State Legislatures, 2024). This fragmentation creates persistent gaps in standards, accountability, and institutional capacity, highlighting the need for harmonized governance and operational guidance (Chandra et al., 2024; Future of Privacy Forum, 2024; UK Department for Science, Innovation and Technology, 2026; UNESCO, 2025a).

2.5 Ethics, Rights and Public Interest

Deepfake governance raises important ethical and human rights considerations. While detection and labeling improve transparency, they also introduce risks associated with false positives, privacy, demographic bias, and potential misuse for censorship (OHCHR, 2024; Romero-Moreno, 2024; UNESCO & OHCHR, 2025; Council of Europe, 2025; Future of Free Speech, 2025). Detection systems may also perform inconsistently across demographic groups and languages because of limitations in training data (International AI Safety Report, 2025; Ma et al., 2025; Wen et al., 2025).

Accordingly, governance frameworks should incorporate transparency, independent oversight, contestability, and regular auditing to ensure fairness and accountability while protecting freedom of expression and public trust (OHCHR, 2024; UNESCO & OHCHR, 2025; Council of Europe, 2025).

2.6 Research gap and Implications for the Proposed Framework

The reviewed literature shows that deepfakes are addressed through several partially connected approaches: technical detection, provenance infrastructure, labelling, platform governance, regulation and media literacy. Communication research emphasises public trust, audience interpretation and institutional legitimacy (Ahmed et al., 2025; Raemy et al., 2025; UNESCO, 2025a). Cybersecurity and computer vision research focus on detection accuracy, robustness and forensic methods (Pei et al., 2026; Prabhu et al., 2026; Raza et al., 2026). Policy research examines transparency, accountability and regulatory fragmentation (European Union, 2024; Future of Privacy Forum, 2024; OECD, 2024). Ethical and rights-based scholarship highlights risks of overreach, bias, privacy intrusion and censorship (OHCHR, 2024; Romero-Moreno, 2024; Council of Europe, 2025).

The main research gap lies in the lack of an integrated framework that connects these domains into practical governance, policy and detection guidelines for communication ecosystems. Detection tools need to be embedded in institutional verification workflows. Labelling mechanisms need to connect machine-readable provenance with user-facing communication. Regulation needs to be translated into operational requirements for platforms, public agencies, newsrooms and other verification actors. Media literacy needs to complement, rather than replace, institutional responsibility (Chandra et al., 2024; OECD, 2024; UNESCO, 2025a).

Therefore, this research proposes a policy-oriented multimodal framework. It is based on three pillars: standardized disclosure and labelling, institutional capacity-building and collaborative governance. The first pillar addresses the need for interoperable human-readable and machine-readable transparency mechanisms. The second pillar focuses on strengthening verification capacity among journalists, public institutions, platforms and civil society actors. The third pillar emphasizes cooperation among public agencies, technology providers, academic researchers, fact-checkers, and media organizations.

3. Research Methodology – Design Science Research

The Design Science Research (DSR) methodology guided the first three phases of this study: problem explication, requirements definition, and artifact design. The problem explication phase identified the limitations of existing deepfake detectors, which typically rely on a single modality (e.g., facial, audio, or metadata analysis) or a single detection algorithm. As generative AI produces increasingly realistic manipulated and fully synthetic media, these approaches have become less effective, motivating the need for a multimodal detection framework.

During requirements definition, the framework was designed to support multimodal feature extraction, ensemble-based image forensics, feature fusion, explainable AI, and confidence-based decision support for both video and still images. The resulting artifact, shown in Figure 1, processes video through facial, audio, and metadata analysis while routing still images and extracted video frames through an ensemble of complementary image-forensics detectors, including artifact, diffusion, semantic, and generative-model fingerprint detectors (Li et al., 2025; Xu et al., 2025).

The outputs are fused into high-dimensional embeddings and combined through calibrated ensemble weighting to generate a confidence score classifying media as authentic, AI-manipulated, or AI-generated. An explainability layer provides human-readable rationales using artifact localization and large vision-language models. Rather than making decisions, these models explain the ensemble's findings, while the calibrated ensemble remains the primary decision engine (Wen et al., 2025; Yu et al., 2025). By integrating visual, audio, forensic, and contextual evidence, the proposed framework provides a more robust and extensible approach than traditional unimodal detection systems, allowing new detection algorithms to be incorporated as generative AI technologies continue to evolve.

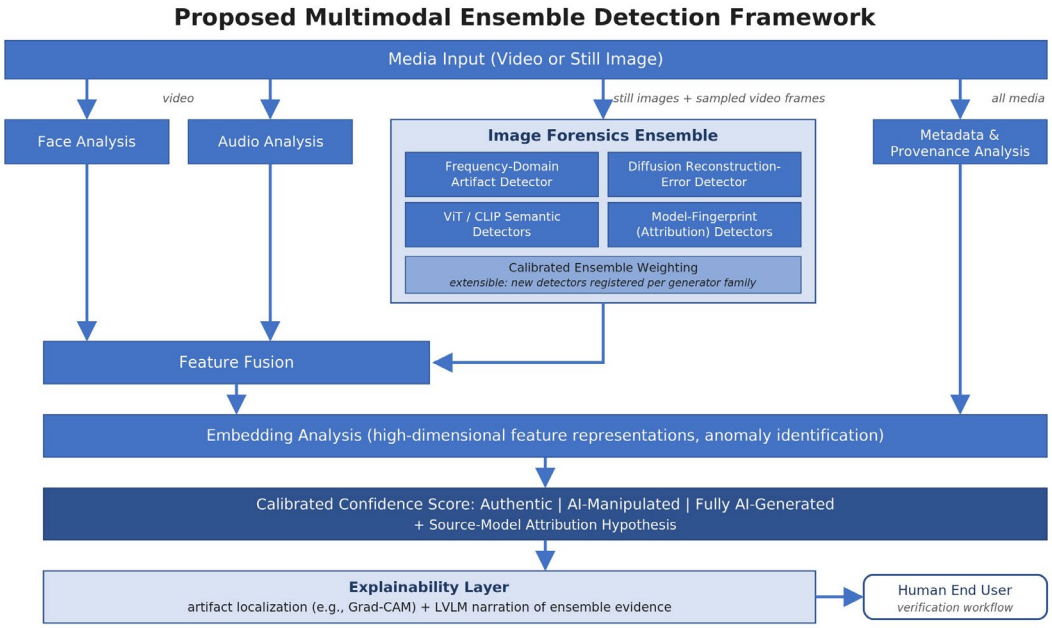

Figure 1; Proposed Multimodal Detection Framework

To evaluate the framework, we propose a structured design of experiments (DOE) in Figure 2. Independent variables include detection methods (individual detectors versus ensemble configurations), media types (video, AI-manipulated images, and fully generated images), generative model families, dataset sources, quality levels, and attack sophistication. Dependent variables include accuracy, precision, recall, F1 score, false positive rates, false negative rates, and processing time. By systematically varying these factors, we can determine which approaches perform best under different conditions and identify limitations that require further research.

Design of Experiments (DOE)

Independent Variables	
Factor	**Levels**
Detection Method	Individual detectors vs. ensemble configurations *(incl. model-fingerprint / attribution detectors)*
Media Type	Video · AI-manipulated image · fully AI-generated image (text-to-image)
Generative Model Family	GAN · diffusion · frontier text-to-image (native multimodal)
Dataset Source	Academic benchmark · in-the-wild · internally generated (experiment corpus)
Quality Level	High · medium · low (compression, downscaling, re-encoding)
Attack Sophistication	Naïve · post-processed · adversarial (detector-evading)

Dependent Variables
Accuracy
Precision
Recall
F1 Score
False Positive Rate
False Negative Rate
Processing Time

Figure 2: DOE for Multimodal Detection

4. Integrated Deepfake Detection and Governance Framework

Neither technical detection nor regulation alone is sufficient to address the challenges posed by deepfakes. The proposed framework integrates the multimodal detection artifact with transparency mechanisms and governance practices into a three-layer architecture consisting of detection and verification, transparency and disclosure, and governance and capacity building. Automated detection serves as decision support rather than a final determination of authenticity.

4.1 Detection and Verification

The framework builds on the multimodal ensemble described in Section 3, using complementary detectors to overcome the limitations of single-model approaches and adapt to emerging generative AI techniques (Chandra

et al., 2026; Li et al., 2025). Both video and still images are analyzed using facial, audio, metadata, and image-forensic evidence, including generative-model fingerprint detection (Xu et al., 2025).

Detection results are expressed as calibrated confidence scores identifying content as authentic, AI-manipulated, or AI-generated, accompanied by source-model attribution and explainable rationales (Yu et al., 2025). Confidence thresholds should be determined by the deploying organization because acceptable risk varies across newsrooms, platforms, and public agencies. Mid-confidence results should undergo human review to reduce false positives and false negatives, while provenance information such as C2PA content credentials should be prioritized whenever available (Brennan Center for Justice, 2024; Chandra et al., 2024; Coalition for Content Provenance and Authenticity, 2025).

4.2 Transparency and Disclosure

The framework leverages transparency mechanisms defined in the EU AI Act, California SB-942, and C2PA by treating metadata, provenance, and watermarking as supporting evidence rather than definitive proof (California Legislature, 2024; European Union, 2024). Verification results are communicated through standardized labels, confidence scores, and explainable rationales instead of binary authenticity claims. This approach improves transparency while enabling regulators and platforms to assess compliance with disclosure requirements. Together with governance and institutional capacity building, these mechanisms strengthen accountability and improve trust in synthetic media verification.

4.3 Governance and Capacity Building

The third layer addresses the institutional gaps identified in Section 2.4. The capacity gap is the most immediate: most newsrooms, public agencies, and civil society organizations cannot maintain a state-of-the-art detection ensemble on their own (UK Department for Science, Innovation and Technology, 2026; UNESCO, 2025a). The framework therefore anticipates a consortium model in which public agencies, technology providers, academic researchers, fact-checkers, and media organizations jointly maintain the detector registry, contribute evaluation data, and share the cost of tracking new generator families, thereby operationalizing the collaborative governance pillar. The standards gap is addressed by anchoring disclosure formats in existing instruments, including the EU AI Act, SB-942, and C2PA, rather than inventing new ones. The accountability gap is addressed through

5. Discussion: Governance and Capacity Building

The proposed framework demonstrates that effective deepfake mitigation depends as much on institutional governance as on technical capability. Because most organizations lack the resources to independently maintain advanced detection systems, a collaborative governance model is necessary. Public agencies, technology providers, academic researchers, fact-checkers, and media organizations should collectively maintain detection capabilities, share evaluation data, and respond to emerging generative AI techniques. Leveraging established standards such as the EU AI Act, California SB-942, and C2PA promotes consistent approaches to disclosure, provenance, and transparency rather than creating competing frameworks.

The findings also highlight the importance of accountability and public trust. Organizations adopting deepfake detection should establish transparent verification procedures, publish confidence thresholds, document decisions, and provide mechanisms for review and appeal. Because detection accuracy may vary across demographic groups, languages, and media types, continuous auditing and public reporting are essential to ensure fairness and reliability. Likewise, media literacy should complement institutional verification by helping users critically evaluate digital content without shifting responsibility away from organizations.

Taken together, the three layers detection, transparency, and governance operate as an integrated system. Detection provides technical evidence, transparency communicates that evidence, and governance establishes the policies and institutional capacity needed to sustain trust. As generative AI continues to evolve, the framework should be viewed as an adaptive governance model that requires ongoing evaluation, refinement, and validation through the proposed Design of Experiments. This combination of technical innovation and collaborative governance represents the paper's primary contribution to strengthening information integrity in modern communication ecosystems.

6. Conclusion

The rapid evolution of generative artificial intelligence has fundamentally altered the challenge of maintaining trust in digital communication. This study demonstrated that effective deepfake mitigation requires more than

improvements in detection accuracy; it demands an integrated approach that combines multimodal technical detection, transparency mechanisms, institutional governance, and public policy. Using the first three phases of Design Science Research, this paper proposed a multimodal detection artifact and embedded it within a broader governance framework that aligns technical verification with emerging regulatory requirements and organizational decision-making.

The study addressed four research questions. RQ1 identified deepfakes as a systemic threat to communication ecosystems by undermining public trust and information integrity. RQ2 demonstrated that current mitigation efforts are converging toward multimodal detection, provenance tracking, standardized labeling, media literacy, and collaborative governance. RQ3 revealed persistent shortcomings in transparency, accountability, institutional capacity, and cross-sector coordination, while RQ4 showed that although the EU AI Act and California AI Transparency Act represent important advances, global governance remains fragmented and lacks interoperable implementation mechanisms. Together, these findings support the proposed framework as a practical model for integrating technical, organizational, and policy responses to synthetic media.

Several limitations should be acknowledged. This research is conceptual and design-oriented; therefore, the proposed framework has not yet undergone empirical validation using operational datasets or real-world deployment. The rapidly changing capabilities of generative AI also mean that detection performance and governance requirements will continue to evolve, potentially reducing the long-term effectiveness of individual detection methods. Furthermore, regulatory analysis focused primarily on the European Union and California, limiting broader international comparisons.

Future research should implement and experimentally evaluate the proposed framework using the Design of Experiments presented in this study, comparing ensemble and single-model detection approaches across diverse datasets, media formats, and adversarial conditions. Additional work should examine explainable AI, fairness across demographic groups, human-in-the-loop verification workflows, and cross-jurisdictional governance models. Ultimately, maintaining trust in communication ecosystems will require continuous collaboration among researchers, policymakers, technology providers, journalists, and public institutions as synthetic media technologies continue to advance.

Ethics Declaration: This research did not involve human participants, human subjects data, or experimental procedures requiring institutional ethical approval. Accordingly, ethical clearance was not required for this study.

AI Declaration: AI-assisted writing tools were used solely to improve grammar, spelling, readability, and language clarity. No AI-generated content was used to develop the research findings, analysis, or conclusions. The authors are fully responsible for the content of this manuscript.

References

Ahmed, S., Masood, M., Bee, A. W. T., & Ichikawa, K. (2025). False failures, real distrust: The impact of an infrastructure failure deepfake on government trust. *Frontiers in Psychology, 16*, Article 1574840. https://doi.org/10.3389/fpsyg.2025.1574840

Al Aghbari, A., Aman, W., Tarhini, A., & Iqbal, W. (2026). Believing the unreal: A systematic review in mapping human responses to deepfakes through empirical evidence. *Social Sciences & Humanities Open, 13*, Article 102772. https://doi.org/10.1016/j.ssaho.2026.102772

Ax, J., & Coster, H. (2026, March 28). AI deepfakes blur reality in 2026 US midterm campaigns. *Reuters*. https://www.reuters.com/business/media-telecom/ai-deepfakes-blur-reality-2026-us-midterm-campaigns-2026-03-28/

Bharati, N., et al. (2025). Explainable deepfake detection: A multi-model framework. *Decision Analytics Journal*. https://www.sciencedirect.com/science/article/pii/S2666827025002026

Brennan Center for Justice. (2024). *Deepfakes, elections, and the shrinking liar's dividend*. https://www.brennancenter.org/our-work/research-reports/deepfakes-elections-and-shrinking-liars-dividend

California Legislature. (2024). *SB-942 California AI Transparency Act*. California Legislative Information. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202320240SB942

Canadian AI Safety Institute. (2025). *Research agenda on risks from synthetic content*. Innovation, Science and Economic Development Canada. https://ised-isde.canada.ca/site/ised/en/canadian-artificial-intelligence-safety-institute/research-agenda-risks-synthetic-content

Cao, Z., Tu, W., Xiao, Y., Deng, W., Chen, W., Lin, L., & Wei, P. (2026). Where detectors fail: Probing generative space for generalizable AI-generated image detection. arXiv. https://arxiv.org/abs/2605.24906

Chandra, B., Dunietz, J., & Roberts, H. (2024). *Reducing risks posed by synthetic content: An overview of technical approaches to digital content transparency* (NIST AI 100-4). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-4

Chandra, N. A., et al. (2026). Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). https://arxiv.org/abs/2503.02857

CIVICUS. (2025). *Future-proofing elections against deepfake disinformation*. https://www.civicus.org/documents/ddi/final-deepfakes-elections-report-civicus-ddi-research-2025.pdf

Coalition for Content Provenance and Authenticity. (2025). *C2PA technical specification*. https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html

Content Credentials. (2025). *Content Credentials*. https://contentcredentials.org/

Council of Europe. (2025). *Guidance note on the implications of generative AI-based systems for freedom of expression*. Council of Europe. https://rm.coe.int/msi-ai-2025-10-draft-guidance-note-on-the-implications-of-generative-a/1680b68c48

Diel, A., Lalgi, T., Mellis, F. S., Teufel, M., & Bäuerle, A. (2025). The harm of deepfakes: A scoping review of deepfakes' negative effects on human mind and behavior. *AI & Society*. https://doi.org/10.1007/s00146-025-02774-0

European Commission. (2025a). *Code of Practice on marking and labelling of AI-generated content*. Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content

European Commission. (2025b, December 17). *Commission publishes first draft of Code of Practice on marking and labelling of AI-generated content*. Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/news/commission-publishes-first-draft-code-practice-marking-and-labelling-ai-generated-content

European Commission. (2026, March 5). *Commission publishes second draft of Code of Practice on marking and labelling of AI-generated content*. Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/library/commission-publishes-second-draft-code-practice-marking-and-labelling-ai-generated-content

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. *Official Journal of the European Union*. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Future of Free Speech. (2025). *"That violates my policies": AI laws, chatbots, and the future of expression*. https://futurefreespeech.org/wp-content/uploads/2025/10/AI-Report-2025-Full-Report.pdf

Future of Privacy Forum. (2024). *Synthetic content report*. https://fpf.org/wp-content/uploads/2024/10/Synthetic_Content_Report_October_2024.pdf

Hastuti, H., Maulana, H. F., Lawelai, H. L., & Suherman, A. S. (2025). Algorithmic influence and media legitimacy: A systematic review of social media's impact on news production. *Frontiers in Communication, 10*, Article 1667471. https://doi.org/10.3389/fcomm.2025.1667471

Hynek, N., Gavurova, B., & Kubak, M. (2025). Risks and benefits of artificial intelligence deepfakes: Systematic review and comparison of public attitudes in seven European countries. *Journal of Innovation & Knowledge, 10*(5), Article 100782. https://doi.org/10.1016/j.jik.2025.100782

International AI Safety Report. (2025). *International AI safety report 2025*. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025

International AI Safety Report. (2026). International AI safety report 2026. arXiv. https://arxiv.org/abs/2602.21012

Iufereva, A. (2024). How is fake news spread? An analysis of the dissemination process: Actors, channels, and motives. *International Conference on AI Research, 4*(1), 456–462. https://doi.org/10.34190/icair.4.1.3054

Li, F., & Yang, Y. (2024). Impact of artificial intelligence–generated content labels on perceived accuracy, message credibility, and sharing intentions for misinformation: Web-based randomized controlled experiment. *JMIR Formative Research, 8*, Article e60024. https://doi.org/10.2196/60024

Li, Z., Yan, J., He, Z., Zeng, K., Jiang, W., Xiong, L., & Fu, Z. (2025). Is artificial intelligence generated image detection a solved problem? In Advances in Neural Information Processing Systems (NeurIPS 2025 Datasets and Benchmarks Track). https://arxiv.org/abs/2505.12335

Ma, Y., Zhang, Y., Chen, X., & Liu, H. (2025). Linguistic features of AI mis/disinformation and the implications for detection. *Nature Communications, 16*, Article 10545. https://doi.org/10.1038/s41467-025-67145-1

National Conference of State Legislatures. (2024). *Deceptive audio or visual media ("deepfakes") 2024 legislation*. https://www.ncsl.org/technology-and-communication/deceptive-audio-or-visual-media-deepfakes-2024-legislation

Nizami, A. H. (2025). Anti-crisis communication strategies in the era of deepfakes: Protecting reputation and restoring trust. *The American Journal of Management and Economics Innovations, 7*(11), 26–33. https://doi.org/10.37547/tajmei/Volume07Issue11-04

OECD. (2024). *Facts not fakes: Tackling disinformation, strengthening information integrity*. OECD Publishing. https://doi.org/10.1787/d909ff7a-en

Office of the United Nations High Commissioner for Human Rights. (2024). *Taxonomy of human rights risks connected to generative AI: Supplement to B-Tech's foundational paper on the responsible development and deployment of generative AI*. https://www.ohchr.org/sites/default/files/documents/issues/business/b-tech/taxonomy-GenAI-Human-Rights-Harms.pdf

Pei, G., et al. (2026). Deepfake generation and detection: A benchmark and survey. *ACM Computing Surveys*. https://doi.org/10.1145/3801962

Prabhu, D., et al. (2026). A survey on multimodal deepfake detection system. In *Proceedings of the International Conference on Artificial Intelligence, Smart Data Analytics and Applications*. Atlantis Press. https://www.atlantis-press.com/proceedings/icaisda-25/126022737

Raemy, P., Bendahan Bitton, D., Ötting, H., Hoffmann, C. P., Godulla, A., & Puppis, M. (2025). Deepfakes and journalism: Normative considerations and implications. *Journalism Studies, 26*(14), 1724–1744. https://doi.org/10.1080/1461670X.2025.2547300

Raza, A., et al. (2026). A comprehensive review of deepfake detection techniques. *AI, 7*(2), Article 68. https://www.mdpi.com/2673-2688/7/2/68

Romero-Moreno, F. (2024). Generative AI and deepfakes: A human rights approach to tackling harmful content. *International Review of Law, Computers & Technology, 38*(3), 297–326. https://doi.org/10.1080/13600869.2024.2324540

Romero-Moreno, F. (2025). Deepfake detection in generative AI: A legal framework for responsible innovation. *Computer Law & Security Review, 57*, Article 106107. https://doi.org/10.1016/j.clsr.2025.106107

Salvi, A. (2024). The attack of the clones: Deepfakes and the evolving landscape of disinformation. In N. Kovalčíková (Ed.), *Hacking minds and machines: Foreign interference in the digital era* (Chaillot Paper No. 184, pp. 32–40). European Union Institute for Security Studies. https://www.iss.europa.eu/sites/default/files/EUISSFiles/CP_184.pdf

Somoray, K., Miller, D. J., & Holmes, N. P. (2025). Human performance in deepfake detection: A systematic review. *Human Behavior and Emerging Technologies*, Article 1833228. https://doi.org/10.1155/hbe2/1833228

UK Department for Science, Innovation and Technology. (2026). *Deepfake detection technology*. GOV.UK. https://www.gov.uk/government/publications/deepfake-detection-technology/deepfake-detection-technology

UNESCO. (2025a). *Deepfakes and the crisis of knowing*. https://www.unesco.org/en/articles/deepfakes-and-crisis-knowing

UNESCO. (2025b). *Freedom of expression, artificial intelligence and elections*. https://unesdoc.unesco.org/ark:/48223/pf0000393473

UNESCO, & Office of the United Nations High Commissioner for Human Rights. (2025). *Protecting critical voices: Guidance for human rights impact assessment on digital platforms*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000396797

Wen, S., Ye, J., Feng, P., Kang, H., Wen, Z., Chen, Y., Wu, J., Wu, W., He, C., & Li, W. (2025). Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. arXiv. https://arxiv.org/abs/2503.14905

Wen, Y., Park, J., Li, M., & Jain, A. K. (2025). AI-Face: A million-scale demographically annotated AI-generated face dataset and fairness benchmark. *arXiv*. https://arxiv.org/abs/2406.00783

Xu, H., Liu, C., Zhu, C., Wang, M., Qu, Y., & Gao, L. (2025). Causal fingerprints of AI generative models. arXiv. https://arxiv.org/abs/2509.15406

Yu, P., Fei, J., Gao, H., Feng, X., Xia, Z., & Chang, C. H. (2025). Unlocking the capabilities of large vision-language models for generalizable and explainable deepfake detection. Proceedings of the 42nd International Conference on Machine Learning, PMLR 267, 72925–72943. https://proceedings.mlr.press/v267/yu25d.html