E-B Environment - Behaviour Proceedings Journal

AicE-Bs2026Sokcho

e-IPH e-International Publishing House Ltd., United Kingdom

https://www.amerabra.org/index.php/aice-bs2026sokcho-04-06-jun-2026/

14th Asia-Pacific International Conference on Environment-Behaviour Studies

Kyungdong University Global (KDU Global),Sokcho (Goseong), South Korea, 04-06 Jun 2026

Deep Fake Technology in Media: A literature review

Mukiibi Moses, Lyimo Johnson Samwel, Gombe Mikael Tengemano, Ahmed Abdulhakim Al-Absi, Baseem Al Athwari*

Department of Smart Computing, Kyungdong University, 46 4 gil, Bongpo, Gosung, Gangwon-do 24764, Korea

Email of ALL Authors:
Mosesmukiibi21@gmail.com, johnsonmoshy6@gmail.com , michaelgombe46@gmail.com, absiahmed@kduniv.ac.kr, baseem_cs@kduniv.ac.kr
*Corresponding author: baseem_cs@kduniv.ac.kr
Tel of 1st author only:+8210-3469-4991

Abstract

With rapid advances in artificial intelligence (AI), deepfake technology can create highly realistic visual and audio media using deep learning and generative adversarial networks (GANs). While beneficial in entertainment, education, and media production, it raises ethical, social, and security concerns. This report outlines its evolution, technical mechanisms, and growing role in misinformation and disinformation. It examines misuse in identity theft, political manipulation, propaganda, and defamation, and their impact on trust. It also reviews countermeasures such as AI detection, watermarking, regulation, and awareness. A cross-disciplinary approach is essential to balance innovation with safeguards and ensure responsible use.

Keywords: Deepfake technology; Artificial intelligence; Identity theft; AI ethics;

DOI:

1.0 Introduction

1.1 Background of Deepfake Technology in Media

The rapid advancement of artificial intelligence (AI) and deep learning has fundamentally transformed modern media, enabling unprecedented capabilities in the creation, manipulation, and dissemination of digital content. Among these developments, deepfake technology has emerged as one of the most powerful and controversial innovations. Deepfakes refer to synthetically generated or manipulated media in which a person's likeness, voice, or actions are realistically altered using advanced machine learning models, particularly Generative Adversarial Networks (GANs), diffusion models, and neural rendering systems (Goodfellow et al., 2014; Mirsky & Lee, 2021).

In media industries, deepfake technology has introduced valuable applications in film production, entertainment, journalism, education, and virtual reality. It has been used for visual effects enhancement, digital resurrection of actors, multilingual dubbing, personalized advertising, and immersive simulations. These applications demonstrate the constructive potential of deepfake systems when deployed ethically and responsibly (Verdoliva, 2020).

eISSN: 2398-4287 © 2026. The Authors. Published for AMER by e-International Publishing House, Ltd., UK. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). Peer–review under responsibility of AMER (Association of Malaysian Environment-Behaviour Researchers).
DOI:

1.2 Problem Statement and Significance

Despite its innovative potential, deepfake technology presents significant threats to media integrity, information authenticity, and public trust. Deepfakes have increasingly been weaponized for misinformation, disinformation, political propaganda, identity theft, financial fraud, defamation, and non-consensual explicit content creation. In media ecosystems, the spread of highly convincing synthetic content undermines journalism credibility, destabilizes political discourse, and contributes to declining confidence in digital information systems (Güera & Delp, 2018).

The rapid sophistication of deepfake generation techniques has outpaced many detection and regulatory systems, creating an ongoing technological arms race between synthetic media creators and defenders. As a result, understanding deepfake technology within media contexts requires interdisciplinary analysis that incorporates technological, ethical, legal, and societal dimensions. This report therefore examines the evolution, applications, threats, and mitigation strategies associated with deepfake technology in media, with particular focus on its implications for trust, security, and governance.

2.0 Literature Review

2.1 Evolution of Deepfake Technology and Media Applications

Deepfake technology has evolved rapidly from early GAN-based systems to more sophisticated diffusion and multimodal generative models. Goodfellow et al. (2014) introduced GANs as the foundational architecture for synthetic media generation, enabling machines to create increasingly realistic visual outputs. Subsequent advances have significantly improved the quality, accessibility, and scalability of manipulated media.

Recent studies indicate that diffusion models and large-scale generative AI systems now produce synthetic content with greater realism than earlier GAN-based systems, making deepfakes more difficult to detect and more impactful within media environments (Edwards et al., 2024; Gong & Li, 2024). These technologies are increasingly integrated into entertainment production, advertising, gaming, and digital storytelling, where they offer creative and commercial benefits.

However, the democratization of deepfake tools through publicly available software and mobile applications has also expanded opportunities for misuse. The literature consistently highlights that as deepfake generation becomes more sophisticated, the risks to media authenticity increase proportionally (Singh et al., 2025; Ming, 2025; Aribe, 2025).

Table 1. Evolution of deepfake generation technologies and key milestones

Period	Technology / Model	Key Capability	Reference
2014–2016	Generative Adversarial Networks (GANs)	Foundational synthetic image generation	Goodfellow et al. (2014)
2017–2019	Face-swapping autoencoders; DeepFaceLab	Realistic video face replacement; public tool availability	Mirsky & Lee (2021)
2020–2022	Neural Rendering; StyleGAN; multimodal synthesis	High-fidelity audio-visual deepfakes; media forensics challenges	Verdoliva (2020)
2023–Present	Diffusion Models; large-scale multimodal generative AI	Superior realism; scalable misuse; harder detection	Edwards et al. (2024); Gong & Li (2024)

(Source: Compiled from Goodfellow et al., 2014; Mirsky & Lee, 2021; Verdoliva, 2020; Edwards et al., 2024; Gong & Li, 2024)

2.2 Detection Technologies and Technical Limitations

A substantial body of research has focused on developing detection systems to counter deepfake misuse. Detection methods commonly utilize convolutional neural networks (CNNs), recurrent neural networks (RNNs), transformers, and hybrid AI systems to identify inconsistencies in manipulated media (Abdulhamed & Hashim, 2024; Khan & Dang-Nguyen, 2023; Rössler et al., 2019; Dolhansky et al., 2020).

Although these systems demonstrate strong performance in controlled laboratory settings, recent studies reveal major limitations in real-world deployment. Lu and Ebrahimi (2024) found that detection models often perform poorly when faced with compressed, noisy, or platform-modified content commonly found on social media platforms. Similarly, Chandra et al. (2025) emphasize that benchmark datasets frequently fail to represent the diversity and complexity of real-world deepfakes, limiting model generalizability. The literature also demonstrates a significant shift from unimodal to multimodal detection strategies. Researchers increasingly advocate

combining visual, audio, textual, and behavioral indicators to address emerging multimodal deepfakes (Liu et al., 2024; Khan & Dai, 2021). While promising, such systems remain underdeveloped and computationally demanding.

Table 2. Comparison of deepfake detection approaches and limitations

Detection Method	Core Approach	Strengths	Key Limitation
CNN-based detection	Visual artifact analysis in image/video frames	High accuracy in lab settings	Poor on compressed/social media content (Lu & Ebrahimi, 2024)
RNN / temporal analysis	Sequential frame inconsistency detection	Captures temporal cues in video	Limited generalizability across novel generation methods (Güera & Delp, 2018)
Transformer-based models	Attention mechanisms over facial features	State-of-the-art accuracy; robust feature learning	Computationally demanding; dataset bias issues (Abdulhamed & Hashim, 2024)
Multimodal detection	Combined visual, audio, and textual cues	Most resilient to emerging deepfake formats	Underdeveloped; lacks standardized benchmarks (Liu et al., 2024; Chandra et al., 2025)

(Source: Compiled from Güera & Delp, 2018; Abdulhamed & Hashim, 2024; Khan & Dang-Nguyen, 2023; Liu et al., 2024; Lu & Ebrahimi, 2024; Chandra et al., 2025)

2.3 Ethical, Social, and Political Impacts on Media

Beyond technical concerns, deepfake technology presents serious ethical and societal challenges for media systems. Scholars widely recognize deepfakes as a growing threat to democratic institutions, journalism credibility, and digital trust. Their use in fake political speeches, election interference, propaganda, and misinformation campaigns has amplified concerns regarding media manipulation and public opinion distortion (Singh et al., 2025; Pawelec, 2022; Vaccari & Chadwick, 2020).

Additionally, deepfakes have facilitated harmful practices such as revenge pornography, identity theft, and financial scams, raising urgent questions surrounding privacy, consent, and legal accountability. These abuses contribute to what some scholars describe as a "liar's dividend," where the existence of deepfakes allows genuine media to be dismissed as fabricated, further eroding trust in authentic journalism and visual evidence (Twomey et al., 2023; Vaccari & Chadwick, 2020).

The broader literature emphasizes that deepfake technology is not solely a technical challenge but a multidisciplinary governance issue requiring collaboration between policymakers, media institutions, AI developers, and educators.

2.4 Research Gaps

Despite significant progress, several major gaps remain in current deepfake research:

- Limited real-world effectiveness of detection systems across diverse platforms and formats.
- Poor cross-domain generalization against newly emerging generation techniques.
- Underdeveloped multimodal detection frameworks.
- Insufficient integration of media ethics, policy, and legal scholarship.
- Lack of standardized global regulatory frameworks for deepfake governance

These gaps suggest that while technological solutions are essential, sustainable responses to deepfake threats in media require broader legal, educational, and institutional interventions.

2.5 Relevance to This Report

The reviewed literature supports this report's focus on deepfake technology as both a transformative and destabilizing force within modern media. It highlights the rapid pace of technological advancement, the vulnerabilities of current detection systems, and the

urgent ethical and governance challenges posed by synthetic media. Consequently, this report adopts a multidisciplinary perspective to assess how media industries can maximize deepfake benefits while minimizing associated harms.

3.0 Methodology

3.1 Research Design

This study adopts a qualitative literature review methodology to critically examine the development, applications, risks, and countermeasures associated with deepfake technology in media. A literature review approach is appropriate because it enables the synthesis of existing scholarly knowledge across multiple disciplines, including artificial intelligence, media studies, cybersecurity, ethics, and digital governance. By analyzing secondary sources, this report identifies key technological advancements, recurring ethical concerns, and emerging policy responses related to deepfake technology.

The research design is primarily descriptive and analytical, focusing on comparing findings from previous academic studies, technical surveys, policy papers, and industry reports. This approach allows for the identification of patterns, research gaps, and interdisciplinary challenges while providing a comprehensive understanding of deepfake technology's influence on media systems.

3.2 Data Collection Methods

Data for this report was collected exclusively from secondary sources, including:Peer-reviewed academic journals

- Conference proceedings
- AI and computer science surveys
- Government and policy reports
- Cybersecurity publications
- Reputable institutional databases

Major academic databases used include Google Scholar, IEEE Xplore, ScienceDirect, SpringerLink, ACM Digital Library, and arXiv. Keywords used during the search process included:

- Deepfake technology
- Deepfake detection
- Synthetic media
- GANs and diffusion models
- AI ethics
- Media misinformation
- Identity theft and digital fraud
- Multimodal deepfake detection

Priority was given to recent studies published between 2020 and 2025 to ensure relevance to the rapidly evolving deepfake landscape, while foundational studies were also included to establish historical and technical context.

3.3 Inclusion and Exclusion Criteria

To maintain academic rigor, sources were selected based on the following inclusion criteria:

- Direct relevance to deepfake technology in media
- Scholarly credibility or institutional authority
- Focus on generation, detection, ethics, regulation, or societal impacts
- Publication in English
- Preference for recent publications

Sources were excluded if they:

- Were outdated without foundational significance
- Lacked scholarly reliability

- Focused solely on unrelated AI applications
- Provided insufficient methodological transparency

Table 3. Inclusion and exclusion criteria for literature selection

Criterion	Inclusion	Exclusion
Topic relevance	Directly addresses deepfake generation, detection, ethics, or regulation	Focused solely on unrelated AI applications
Publication period	2020–2025 (foundational works from 2014 onward also included)	Outdated sources without foundational significance
Source credibility	Peer-reviewed journals, conference proceedings, institutional reports	Sources lacking scholarly reliability or methodological transparency
Language	English-language publications	Non-English works (unless translated abstracts confirmed relevance)

(Source: Authors)

3.4 Data Analysis Approach

The collected literature was analyzed using thematic analysis, allowing studies to be grouped into major categories:

- Evolution of deepfake generation technologies
- Media applications and benefits
- Detection technologies and technical limitations
- Ethical, legal, and social concerns
- Policy responses and countermeasures

This thematic framework facilitated comparative evaluation across studies and supported the identification of recurring trends, contradictions, and unresolved challenges. Through this process, the report assesses both the opportunities and threats posed by deepfake technology while emphasizing the need for multidisciplinary solutions.

3.5 Limitations of the Methodology

Although the literature review approach provides broad academic coverage, it is limited by reliance on existing published materials rather than primary empirical investigation. Additionally, the rapid pace of AI development means some technological advancements may evolve faster than available scholarly literature. Variations in dataset quality, publication standards, and regional policy frameworks may also influence findings.

Despite these limitations, the methodology provides a robust framework for critically evaluating deepfake technology in media and offers a strong basis for identifying future research and policy directions.

4.0 Findings

4.1 Positive Applications of Deepfake Technology in Media

The literature reveals that deepfake technology offers several beneficial applications when used ethically within media industries. In entertainment and film production, deepfakes have enhanced visual effects, enabled digital de-aging, facilitated actor resurrection, and improved multilingual dubbing, reducing production costs while expanding creative possibilities (Edwards et al., 2024). Media organizations have also explored synthetic voice and visual technologies for personalized advertising, virtual presenters, and immersive storytelling experiences (Verdoliva, 2020).

In education and training, deepfake-based simulations can improve engagement through realistic historical recreations, language learning, and virtual instruction (Mirsky & Lee, 2021). Additionally, accessibility improvements such as AI-generated voice synthesis have created opportunities for individuals with speech impairments (Kulangareth et al., 2024). Deepfake technology also holds potential

for scientific education, though it simultaneously poses risks to scientific knowledge integrity (Doss et al., 2023). These findings suggest that deepfake technology itself is not inherently harmful; rather, its societal impact depends largely on context, intent, and governance. Responsible deployment can support innovation, creativity, and digital transformation in media ecosystems.

4.2 Risks, Threats, and Societal Consequences

Despite these benefits, the dominant findings in current literature emphasize the substantial risks associated with malicious deepfake use. Deepfakes have increasingly been weaponized for:

- Political misinformation and election interference
- Fake news dissemination
- Financial fraud and social engineering scams
- Identity theft and impersonation
- Defamation and reputational harm
- Non-consensual explicit content creation

A major consequence identified across studies is the erosion of public trust in digital media. As synthetic content becomes more realistic, distinguishing authentic journalism from manipulated media becomes increasingly difficult, undermining confidence in visual and audio evidence. This challenge contributes to the broader phenomenon of declining trust in media institutions and democratic discourse.

The literature also highlights the "liar's dividend," where genuine media can be dismissed as fake, allowing individuals or institutions to evade accountability. This not only threatens media credibility but also weakens legal and political systems dependent on reliable evidence.

Table 4. Summary of deepfake threat categories, impacts, and affected domains

Threat Category	Description	Affected Domain	Reference
Political misinformation	Fabricated speeches and election interference via synthetic video	Democracy; journalism	Singh et al. (2025)
Identity theft & fraud	Impersonation of individuals for financial scams and reputational harm	Finance; cybersecurity	Güera & Delp (2018); Mirsky & Lee (2021)
Non-consensual explicit content	Fabricated intimate media to harass or defame victims	Privacy; personal safety; law	Singh et al. (2025)
Erosion of media trust ("liar's dividend")	Genuine media dismissed as fake; declining public confidence in information	Media; governance; public discourse	Verdoliva (2020); Mirsky & Lee (2021)

(Source: Compiled from Güera & Delp, 2018; Mirsky & Lee, 2021; Verdoliva, 2020; Singh et al., 2025)

4.3 Detection Challenges and Countermeasure Limitations

Current detection technologies, while increasingly sophisticated, remain insufficient in addressing rapidly evolving deepfake generation systems. AI-based detectors using CNNs, transformers, and multimodal systems perform relatively well in controlled settings but often struggle in real-world environments characterized by social media compression, diverse content formats, and novel generation methods. Furthermore, findings indicate that deepfake generation technology often evolves faster than detection systems, creating an ongoing technological arms race. Existing countermeasures such as watermarking, digital forensics, and content moderation policies offer partial solutions but face implementation, scalability, and enforcement limitations.

Overall, the findings demonstrate that technical solutions alone are unlikely to fully address deepfake threats without complementary legal, ethical, and educational strategies.

5.0 Discussion

5.1 Implications for Media Integrity, Governance, and Society

The findings underscore that deepfake technology represents a significant challenge to the integrity of modern media systems. The ability to convincingly manipulate visual and auditory content threatens the foundational role of media as a trusted source of information. Journalism, political communication, and social media platforms are particularly vulnerable, as deepfakes can rapidly spread misinformation, influence public opinion, and destabilize democratic institutions.

From a governance perspective, current regulatory responses remain fragmented and inconsistent across jurisdictions. Many legal systems lack comprehensive frameworks specifically addressing synthetic media misuse, while social media platforms continue to face difficulties balancing content moderation with freedom of expression (Singh et al., 2025). This regulatory lag creates opportunities for malicious actors to exploit technological vulnerabilities.

The discussion also highlights the ethical complexity surrounding deepfakes. While beneficial uses exist, the potential for abuse raises serious concerns regarding consent, privacy, and accountability. Therefore, deepfake governance must move beyond purely technical approaches to include policy reform, ethical AI development, public awareness, and international cooperation (Romero Moreno, 2024; Pawelec, 2022).

5.2 Future Directions and Strategic Recommendations

Addressing deepfake challenges will require a multidisciplinary and collaborative approach. Future strategies should prioritize:

- Advanced multimodal detection systems capable of adapting to emerging threats (Man & Cho, 2025; Singh & Dhumane, 2025)
- Standardized digital watermarking and authentication protocols
- Stronger international legal and regulatory frameworks
- Enhanced platform accountability for synthetic media moderation (Gao et al., 2026)
- Public digital literacy campaigns to improve misinformation resilience (Hwang & Ryu, 2021)
- Increased collaboration between academia, industry, and policymakers

In addition, media organizations must strengthen verification processes and invest in forensic capabilities to maintain journalistic credibility. Educational initiatives are equally essential, as public awareness remains one of the most effective defenses against synthetic media deception.

Ultimately, the future of deepfake technology in media will depend on society's ability to balance innovation with responsible governance. Without proactive intervention, the risks to trust, security, and democratic integrity may outweigh the technology's creative and commercial benefits.

6.0 Conclusion and Recommendations

6.1 Conclusion

Deepfake technology has emerged as one of the most transformative yet disruptive developments in contemporary media. Powered by advances in artificial intelligence, deep learning, GANs, and diffusion models, deepfakes have introduced substantial opportunities for innovation across entertainment, education, advertising, and digital communication. Their ability to enhance creative production and personalized media experiences demonstrates significant technological promise.

However, the literature overwhelmingly indicates that deepfake technology also presents severe ethical, legal, social, and political risks. Its misuse in misinformation campaigns, political manipulation, identity theft, financial fraud, and non-consensual exploitation poses direct threats to media credibility, democratic stability, and public trust. As synthetic media becomes increasingly realistic and accessible, distinguishing truth from manipulation grows progressively more difficult, creating long-term challenges for journalism, governance, and societal trust.

Current detection systems and policy responses, while improving, remain insufficient to fully address the rapidly evolving nature of deepfake generation. The persistent arms race between creation and detection technologies highlights the limitations of technical solutions alone. Therefore, deepfake governance must be approached as a multidisciplinary issue that requires coordinated efforts across technological, legal, ethical, educational, and institutional domains.

Ultimately, the future impact of deepfake technology in media will depend on society's ability to harness its benefits responsibly while

implementing effective safeguards against abuse. Without proactive governance and sustained collaboration, the threats posed by deepfakes may significantly undermine the integrity of digital media ecosystems.

6.2 Recommendations

To effectively address the opportunities and risks associated with deepfake technology in media, the following recommendations are proposed:

For Policymakers and Regulators:

- Develop comprehensive legal frameworks specifically addressing malicious deepfake creation and distribution.
- Promote international cooperation to establish standardized global governance protocols.
- Strengthen privacy, identity protection, and cybersecurity legislation.
- Encourage mandatory transparency measures for synthetic media content.

For Technology Developers and AI Researchers:

- Invest in adaptive multimodal detection systems capable of addressing future generation models.
- Integrate watermarking, authentication, and provenance-tracking systems into AI-generated content.
- Prioritize ethical AI development principles and misuse prevention mechanisms.
- Expand collaboration with policymakers and media institutions.

For Media Organizations and Digital Platforms:

- Strengthen fact-checking, verification, and forensic analysis systems.
- Improve platform moderation policies for synthetic media.
- Increase transparency in content labeling and user warnings.
- Provide specialized journalist training in deepfake detection.

For Educational Institutions and Society:

- Promote digital literacy and misinformation awareness programs.
- Educate users on identifying manipulated content.
- Encourage responsible media consumption and verification practices.
- Foster broader societal awareness regarding the ethical implications of synthetic media.

6.3 Final Reflection

Deepfake technology represents both a remarkable technological achievement and a profound societal challenge. Its future role in media will depend not solely on innovation, but on humanity's collective capacity to regulate, understand, and ethically govern its use. Balancing progress with protection is essential to preserving trust, authenticity, and security in the digital age.

Acknowledgements

The authors would like to express their sincere appreciation to the Department of Smart Computing at Kyungdong University for providing the academic environment and resources that supported this study. The authors also extend their gratitude to the anonymous reviewers for their valuable comments and constructive suggestions, which helped improve the quality of this manuscript. This research received no external funding.

Paper Contribution to Related Field of Study

This paper contributes to the field of media studies and artificial intelligence by providing a comprehensive and up-to-date literature review on deepfake technology within modern media environments. It synthesizes recent advancements in deepfake generation, detection techniques, and multimodal AI systems while highlighting critical limitations in real-world deployment. The study further contributes by integrating technical analysis with ethical, social, and governance perspectives, offering a multidisciplinary understanding of deepfake impacts on media integrity, public trust, and digital communication. Additionally, the paper identifies key research gaps, including limitations in detection generalizability, lack of standardized regulatory frameworks, and

insufficient multimodal detection approaches.
Finally, this work provides strategic recommendations for policymakers, researchers, media organizations, and educators, contributing to ongoing efforts toward responsible AI deployment and sustainable governance of synthetic media technologies.

References

Abdulhamed, M. A., & Hashim, A. N. (2024). A survey on detecting deep fakes using advanced AI-based approaches. *Iraqi Journal of Science*.Chandra, N. A., et al. (2025). Deepfake-Eval-2024:

A multimodal in-the-wild benchmark of deepfakes. *arXiv Preprint*.
Edwards, P., Nebel, J.-C., Greenhill, D., & Liang, X. (2024).

A review of deepfake techniques: Architecture, detection and datasets. *IEEE Access*.

Goodfellow, I., et al. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* (pp. 2672–2680).

Gong, L. Y., & Li, X. J. (2024). A contemporary survey on deepfake detection: Datasets, algorithms, and challenges. *Electronics, 13*(3).

Güera, D., & Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. In *Proceedings of the IEEE International Workshop on Information Forensics and Security*.

Khan, S. A., & Dang-Nguyen, D.-T. (2023). Deepfake detection: A comparative analysis. *arXiv Preprint*.

Liu, P., Tao, Q., & Zhou, J. T. (2024). Evolving from single-modal to multi-modal facial deepfake detection: A survey. *arXiv Preprint*.

Lu, Y., & Ebrahimi, T. (2024). Assessment framework for deepfake detection in real-world situations. *EURASIP Journal on Image and Video Processing*.

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys, 54*(1), 1–41.

Singh, S., Srivastav, S., Singh, S., & Hussain, S. (2025). Deepfake detection systems: A comprehensive survey of algorithms and techniques.

Verdoliva, L. (2020). Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing, 14*(5), 910–932.

Doss, C., Mondschein, J., Shu, D., Wolfson, T., Kopecky, D., Fitton-Kane, V. A., Bush, L., & Tucker, C. (2023). Deepfakes and scientific knowledge dissemination. *Scientific Reports, 13*, 13439. https://doi.org/10.1038/s41598-023-39944-3

Gao, L., Ahmed, A., Chen, O., Reyl, M., Cheema, Z., Feamster, N., Tan, C., Thomas, K., & Chetty, M. (2026). Governance of AI-generated content: A case study on social media platforms. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. ACM. https://doi.org/10.1145/3772318.3790415

Hwang, Y., & Ryu, J. Y. (2021). Effects of disinformation using deepfake: The protective effect of media literacy education. *Cyberpsychology, Behavior, and Social Networking, 24*(3), 184–192. https://doi.org/10.1089/cyber.2020.0174

Khan, S. A., & Dai, H. (2021). Video transformer for deepfake detection with incremental learning. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21)* (pp. 1821–1829). ACM. https://doi.org/10.1145/3474085.3475332

Kulangareth, N. V., Kaufman, J., Oreskovic, J., & Fossat, Y. (2024). Investigation of deepfake voice detection using speech pause patterns: Algorithm development and validation. *JMIR Biomedical Engineering, 9*, e56245. https://doi.org/10.2196/56245

Man, Q., & Cho, Y.-I. (2025). Exposing face manipulation based on generative adversarial network–transformer and fake frequency noise traces. *Sensors, 25*(5), 1435. https://doi.org/10.3390/s25051435

Ming, L. (2025). A review of deepfake and its detection: From generative adversarial networks to diffusion models. *International Journal of Intelligent Systems, 2025*, 9987535. https://doi.org/10.1155/int/9987535

Moyo, B. V. C., & Obagbuwa, I. C. (2026). An AI-driven conceptual framework for detecting fake news and deepfake content: A systematic review. *Frontiers in Artificial Intelligence, 9*. https://doi.org/10.3389/frai.2026.1737790

Pawelec, M. (2022). Deepfakes and democracy (theory): How synthetic audio-visual media for disinformation and hate speech threaten core democratic functions. *Digital Society, 1*(3), 19. https://doi.org/10.1007/s44206-022-00010-6

Piazza, A., & Voltolini, F. (2023). Designed to abuse? Deepfakes and the non-consensual diffusion of intimate images. *Synthese, 200*(2), 144. https://doi.org/10.1007/s11229-022-04012-2

Romero Moreno, F. (2024). Generative AI and deepfakes: A human rights approach to tackling harmful content. *International Review of Law, Computers & Technology, 38*(3), 219–242. https://doi.org/10.1080/13600869.2024.2324540

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 1–11). IEEE. https://doi.org/10.1109/ICCV.2019.00009

Schiavone, G., & Maulini, G. (2025). Deepfake detection in generative AI: A legal framework proposal to protect human rights. *Computer Law & Security Review, 57*, 106121. https://doi.org/10.1016/j.clsr.2025.106121

Singh, S., & Dhumane, A. (2025). Unmasking digital deceptions: An integrative review of deepfake detection, multimedia forensics, and cybersecurity challenges. *MethodsX, 2025*, 103632. https://doi.org/10.1016/j.mex.2025.103632

Snauffer, A., Nyst, C., & Bursens, P. (2024). Non-consensual synthetic intimate imagery: Prevalence, attitudes, and knowledge in 10 countries. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2402.01721

Twomey, J., Ching, D., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2023). Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. *PLOS ONE, 18*(10), e0291668. https://doi.org/10.1371/journal.pone.0291668

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

Dolhansky, B., et al. (2020). The deepfake detection challenge (DFDC) dataset. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2006.07397

Geng, C. (2023). Comparing deepfake regulatory approaches. *Georgetown Law Technology Review, 7*, 157–184.

Aribe, S. (2025). Generative artificial intelligence and the evolving challenge of deepfake detection: A systematic analysis. *Journal of Sensor and Actuator Networks, 14*(1), 17. https://doi.org/10.3390/jsan14010017

Chukwudi, O., & Osei, K. (2026). A comprehensive review of deepfake detection techniques: From traditional machine learning to advanced deep learning architectures. *AI, 7*(2), 68. https://doi.org/10.3390/ai7020068