

AI innovation vs. disinformation: The dual role of deepfake technology

Temirlan Ramazan^{1,*†}, Ha Jin Hwang^{1,†}, Askar Murat^{1,2,†} and Aidiye Aidarbekov^{1,2,3,†}

¹ Astana IT University, Mangilik El 55/11 010000 Astana, Kazakhstan

² Kazakh AgroTechnical Research University, Zhenis Avenue 62 010011 Astana, Kazakhstan

³ Maqsut Narikbayev University, Korgalzhyn Highway 8 010000 Astana, Kazakhstan

Abstract

Deepfake technology exemplifies the dual-use nature of artificial intelligence: it is both a catalyst for innovation and a threat to trust and justice. This article examines the paradoxical role of deepfakes through a qualitative synthesis of ten peer-reviewed studies published between 2021 and 2025. Using the United Nations Sustainable Development Goals as a framework, the analysis maps opportunities under **SDG 9 (Industry, Innovation, and Infrastructure)** and risks under **SDG 16 (Peace, Justice, and Strong Institutions)**.

The review shows that deepfakes contribute to industrial modernization and cultural production by enabling creative partnerships, educational innovation, and provenance standards that strengthen technological infrastructure. At the same time, they undermine democratic institutions by amplifying disinformation, eroding trust, and facilitating image-based sexual violence. Evidence also reveals an asymmetry of harms: while the benefits of innovation are broadly distributed, the risks disproportionately affect women, minors, and digitally vulnerable populations.

The findings highlight fragmented global governance, with most research focusing on technical detection rather than societal interventions. To address this imbalance, the article proposes a **dual-track governance model**: fostering responsible innovation in creative industries while embedding safeguards through regulation, platform accountability, and digital literacy.

In conclusion, deepfakes represent neither wholly beneficial nor wholly harmful technologies but context-dependent phenomena. Their future trajectory will be determined by the capacity of societies to design governance ecosystems that maximize innovation while protecting dignity, justice, and institutional trust.

Keywords

deepfake technology, synthetic media, AI innovation, disinformation, public trust, SDG 9, and SDG 16¹

1. Introduction

Artificial intelligence (AI) has transformed the creation and circulation of media, with deepfake technology representing one of its most striking and controversial developments. Deepfakes, created through advanced machine learning techniques such as generative adversarial networks (GANs) and autoencoders, produce synthetic images, audio, and video that are often indistinguishable from authentic recordings [1]. They hold great promise for creative industries, education, and accessibility [2], [11], but they also pose unprecedented challenges for disinformation, security, and governance [4], [6], [8].

The significance of deepfakes lies in their dual-use nature. On one side, they enhance industrial and cultural infrastructures, enabling creative partnerships, educational innovations, and new standards for provenance and media verification [1], [2]. These opportunities align closely with Sustainable Development Goal (SDG) 9: Industry, Innovation, and Infrastructure. On the other side, deepfakes are weaponized to spread disinformation, exploit individuals through non-consensual

¹ AI 2025: Workshop on Artificial Intelligence, November 19-20, 2025, Almaty, Kazakhstan

* Corresponding author.

† These authors contributed equally.

✉ temirlan030303@gmail.com (T. Ramazan); hjhwang@astanait.edu.kz (H. J. Hwang); askar.murat@astanait.edu.kz (A. Murat); aidiye.aidarbekov@astanait.edu.kz (A. Aidarbekov);

ORCID 0009-0005-8882-6003 (T. Ramazan); 0000-0001-8752-1906 (H. J. Hwang); 0009-0005-8882-6003 (A. Murat); 0000-0001-7197-1638 (A. Aidarbekov)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

pornography, and undermine trust in institutions [5], [6], [8]. These risks threaten the realization of SDG 16: Peace, Justice, and Strong Institutions.

Deepfakes at the Crossroads: Innovation vs. Disinformation

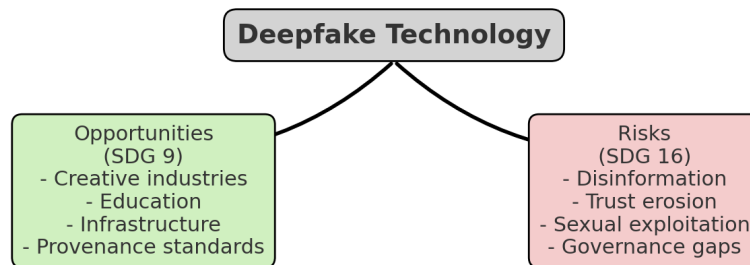


Figure 1: Dual role of deepfake technology in relation to SDG 9 (Innovation) and SDG 16 (Disinformation). Figure generated with the assistance of ChatGPT (OpenAI, 2025).

The scientific value of studying deepfakes lies in their interdisciplinary relevance. Computer science provides the technical basis for generating and detecting synthetic content [7]. Social sciences illuminate how publics perceive and respond to deepfakes [6]. Legal scholarship maps fragmented governance responses [5]. Psychology and communication studies explore their effects on trust, persuasion, and vulnerability [9]. Together, these perspectives highlight both the opportunities of deepfake innovation and the dangers of misuse.

This article aims to synthesize the dual role of deepfakes through a structured literature-based analysis. Drawing on ten scholarly contributions, it organizes findings into four thematic areas: creative and industrial innovation, disinformation and trust erosion, governance and regulation, and asymmetry of harms. By coding evidence through the SDG framework, the study clarifies how deepfakes simultaneously advance innovation and destabilize institutions. The article concludes with a discussion of governance challenges, ethical implications, research gaps, and the proposal of a dual-track strategy for balancing innovation with risk mitigation.

2. Literature review

Creative and Industrial Innovation (SDG 9)

Deepfakes have significant potential to contribute to industrial modernization and creative expression. A foundational review of AI in creative industries highlights applications in content creation, media enhancement, data compression, gaming, and advertising [1]. It is argued that AI is most effective when augmenting, rather than replacing, human creativity [1]. This perspective is reinforced by findings that identify AI-driven art and media as a central research theme in socio-cultural AI studies, with rapid growth in creative applications since 2018 [3].

In education and pedagogy, deepfakes and related generative AI tools are increasingly seen as collaborative partners. A posthuman perspective demonstrates how AI can act as a co-creator in classroom projects and artistic production [11]. These findings suggest that AI can inspire new forms of hybrid creativity while raising ethical questions about authorship and originality [11]. Empirical evidence from creative professionals further shows that trust in AI as a collaborator depends on transparency, ethical safeguards, and the preservation of human agency [9].

Beyond cultural innovation, deepfakes stimulate the development of infrastructural standards. Bibliometric mapping highlights how fake news research has accelerated advances in watermarking, provenance tracking, and detection systems such as the Coalition for Content Provenance and Authenticity (C2PA) [4]. These technical solutions, while designed to mitigate risks, also represent positive innovations in themselves, advancing the infrastructural resilience targeted by SDG 9 [4].

Taken together, these studies demonstrate that deepfakes can enrich cultural production, educational practice, and technological infrastructure, making them powerful enablers of SDG 9 when deployed responsibly.

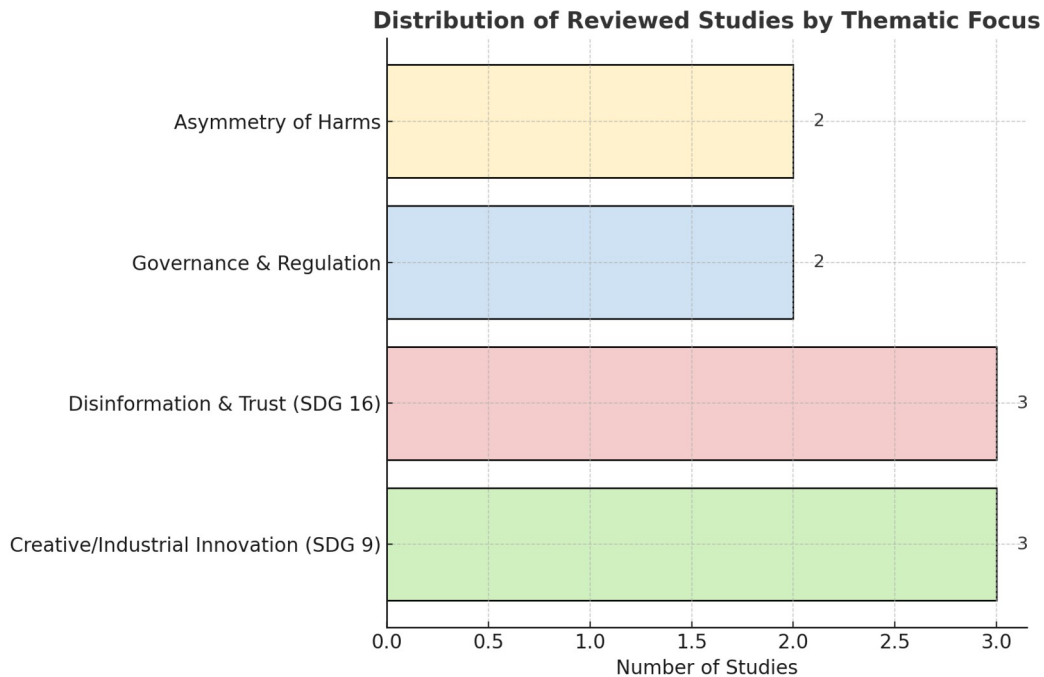


Figure 2: Showing the distribution of reviewed studies by theme (e.g., Innovation, Trust/Disinformation, Governance, Harms). Figure generated with the assistance of ChatGPT (OpenAI, 2025).

Disinformation and Trust Erosion (SDG 16)

A bibliometric study provides quantitative evidence of how fake news and generative AI research align with the Sustainable Development Goals [4]. The analysis shows that 857 publications map to SDG 16 (Peace, Justice, and Strong Institutions), reflecting the widespread concern with disinformation, trust erosion, and institutional legitimacy [4]. At the same time, 214 publications are linked to SDG 9 (Industry, Innovation, and Infrastructure), highlighting the growing focus on provenance standards, detection systems, and resilient information infrastructures [4]. These findings underscore that while health misinformation (SDG 3) dominates globally, the issues most relevant to deepfakes, such as trust in institutions (SDG 16) and technological innovation (SDG 9), are already recognized as major research priorities [4].

This bibliometric evidence strengthens the argument that deepfakes sit at the intersection of innovation and disinformation, directly engaging with both SDG 9 and SDG 16. Building on this perspective, a cross-national survey of public attitudes across seven European countries finds that citizens consistently perceive deepfakes as more harmful than beneficial [6]. The data also reveal generational divides, with younger people more optimistic about creative uses, while older and less digitally literate groups express heightened concern [6].

Recent studies highlight that the implications of deepfake technology extend far beyond technical detection and encompass a broad range of social, ethical, and psychological dimensions. Wazid *et al.* [12] provide one of the most comprehensive analyses of these effects, emphasizing that deepfakes are not merely a cybersecurity challenge but also a significant societal phenomenon that undermines trust, distorts political processes, and compromises individual privacy. Their section “*Societal Impact of Deepfake*” identifies several critical consequences:

- (1) the dissemination of disinformation and the erosion of public trust in media institutions;
- (2) manipulation of political communication and electoral influence;

- (3) privacy and identity theft through synthetic impersonation; and
- (4) the psychological fatigue and skepticism experienced by audiences exposed to fabricated content.

Such observations reinforce the argument that deepfakes directly contribute to trust degradation within digital ecosystems, aligning with the broader framework of information disorder discussed by Raman *et al.* [6] and Daly *et al.* [8].

Integrating Wazid *et al.*'s findings expands the review's scope beyond purely algorithmic detection, demonstrating that the societal dimension of deepfake dissemination has measurable consequences on public confidence and civic stability. This approach justifies a targeted qualitative sample focusing on studies that examine social trust, misinformation, and media ethics, rather than the technical parameters of deepfake detection.

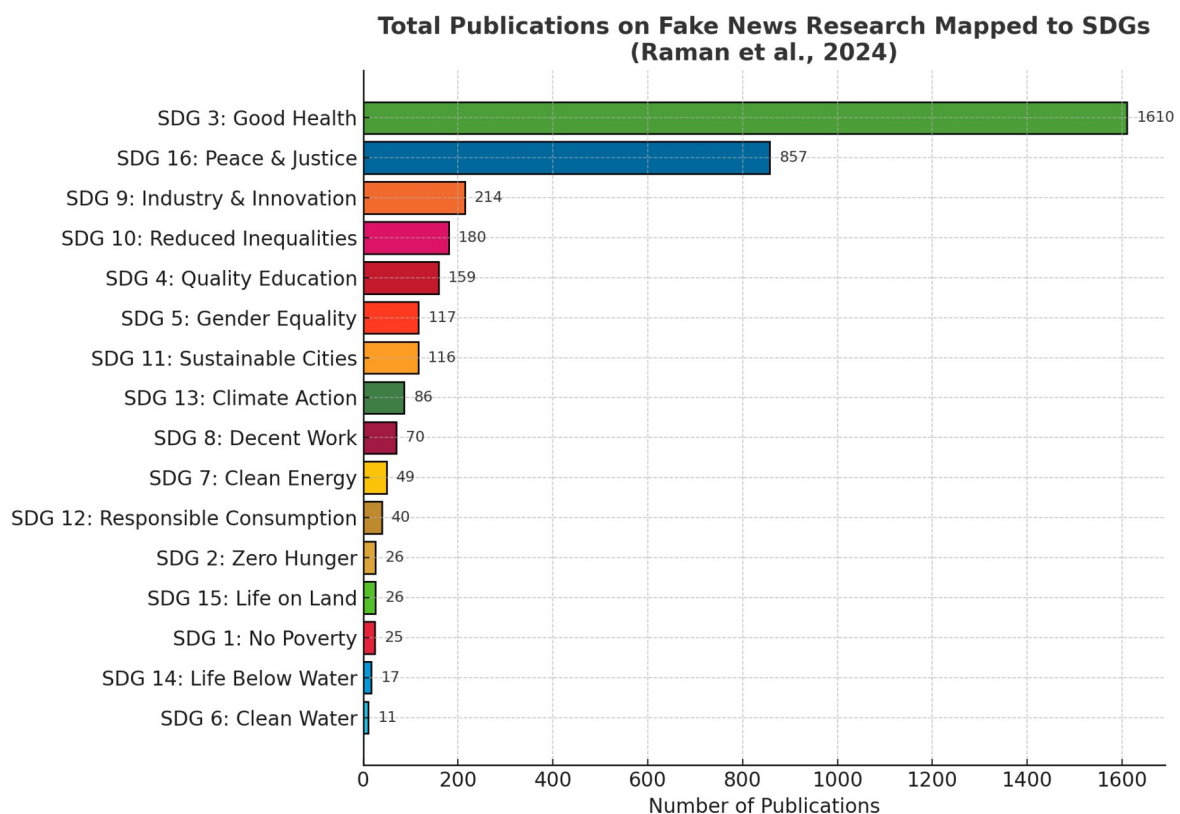


Figure 3: Total publications on fake news research mapped to SDGs, redrawn based on data from Raman et al. (2024). Visualization generated with the assistance of ChatGPT (OpenAI, 2025)

The concept of the liar's dividend has been introduced to describe how deepfakes erode institutional trust not only by deceiving viewers but also by enabling the dismissal of authentic evidence as fabricated [5]. This dynamic undermines journalism, courts, and democratic processes, directly threatening SDG 16's focus on justice and strong institutions [5]. A complementary framework further shows that trust and distrust in deepfakes depend on psychological factors, including the realism of content, credibility of sources, and emotional appeal of narratives [7]. These findings demonstrate how persuasive deepfakes can bypass skepticism, making them potent tools of manipulation [7].

Expanding this perspective, Matli [11] applies the information poverty theory to the domain of synthetic media, demonstrating that deepfakes exacerbate information inequality and asymmetry by disproportionately impacting individuals and communities with limited media literacy or restricted access to reliable information sources. This approach situates deepfake technology within a framework of digital inequality, showing that the manipulation of visual and auditory content not only distorts perception but also amplifies societal divides between those who can critically evaluate

information and those who cannot. Such an understanding aligns with the argument that the threat of deepfakes lies not merely in deception, but in their potential to deepen existing disparities in trust, awareness, and information accessibility across different social groups.

In addition to the growing body of research on deepfake detection, recent scholarship has increasingly examined the ethical and social dimensions of synthetic media. Kaur *et al.* [13] emphasize that deepfakes exacerbate the digital divide by amplifying disparities in media literacy, information access, and critical evaluation skills across different social groups. This unequal exposure leads to varying levels of public trust in online information, contributing to a broader *crisis of authenticity* within digital ecosystems. Their findings complement the theoretical perspectives of Matli [9], who links deepfakes to information poverty, suggesting that both frameworks – *digital divide* and *information inequality* – help explain the erosion of trust in contemporary media environments.

The literature makes clear that deepfakes extend beyond individual deception to systemic trust erosion. By destabilizing confidence in media, politics, and judicial processes, they jeopardize the peace and justice commitments of SDG 16.

Governance and Regulation

Governance responses to deepfakes remain fragmented and reactive. Comparative analysis of major jurisdictions shows that the EU has adopted proactive frameworks such as the AI Act and Digital Services Act, while the US relies on fragmented state-level laws. China enforces upstream algorithm reviews and labeling requirements, and the UK emphasizes platform accountability through the Online Safety Act [5]. These divergent approaches reveal limited global harmonization [5].

Most legal systems also lack explicit definitions of deepfake pornography, leaving victims of image-based abuse without effective legal protection [8]. Cross-national surveys further show that publics demand stronger regulations, though the intensity of this demand varies by country [6]. At the same time, research disproportionately focuses on technical detection, with only a small proportion (about 6 percent) addressing governance, ethics, or regulation [10].

The consensus across these studies is that governance must be multi-layered: technical safeguards such as watermarking and explainable AI, legal frameworks including criminalization of abuse and harmonized international laws, and social interventions like digital literacy and victim-centered justice mechanisms [5], [6], [8], [10].

Asymmetry of Harms

While the benefits of deepfakes are broadly distributed across industries and educators, their harms are highly uneven. Evidence shows that women and minors overwhelmingly bear the burden of deepfake pornography, with nearly 99 percent of victims being female [8]. Older and less digitally literate populations are more vulnerable to disinformation [6], while marginalized groups face greater risks in regions with weaker regulation and digital literacy [10].

This asymmetry underscores that governance must be equitable and inclusive. Without deliberate efforts to protect vulnerable populations, deepfake innovation risks amplifying existing inequalities while disproportionately burdening those least able to defend themselves [6], [8], [10].

3. Methodology

Research Design

This article adopts a qualitative, literature-based research design. Instead of collecting new empirical data, it synthesizes findings from diverse disciplines, such as computer science, law, psychology, media studies, and education, to analyze the dual role of deepfakes. The study is interpretive, integrative, and explicitly framed by the Sustainable Development Goals (SDGs), focusing on SDG 9 (innovation) and SDG 16 (justice).

The guiding research questions were:

1. How do deepfake technologies contribute to industrial innovation and creative industries in line with SDG 9?

2. How do deepfakes threaten peace, justice, and institutional trust in relation to SDG 16?

Corpus Selection

Source Identification

Literature was identified using systematic database searches (Scopus, Web of Science, ScienceDirect) with keywords such as deepfake, generative AI, disinformation, creative industries, trust, governance, SDG. Additional works were identified through expert selection based on relevance to deepfake governance and sustainability. This mix of approaches provides a broad and interdisciplinary base for synthesis.

Although this paper does not constitute a fully systematic review, its scope is intentionally narrow to maintain alignment with the SDG framework. The corpus was restricted to peer-reviewed publications that explicitly link deepfake technologies to SDG 9 (industry, innovation) and SDG 16 (peace, justice, and strong institutions). This deliberate focus ensures conceptual depth and allows for a thematic synthesis centered on sustainability, governance, and trust rather than a broad technical survey. While the corpus remains limited in size, its conceptual alignment with the SDG framework ensures analytical coherence and methodological rigor.

Author(s)	Year	Title	Type of Study	SDG Focus	Inclusion Criteria	Exclusion Criteria
Anantrasirichai & Bull	2021	Artificial intelligence in the creative industries: a review	Review	SDG 9 (Innovation in creative industries)	Peer-reviewed; explicit focus on AI in creative sectors	None – addressed societal/ethical aspects in addition to technical
Feher & Katona	2021	Fifteen shadows of socio-cultural AI: A systematic review and future perspectives	Systematic review	SDG 9 (Socio-cultural innovation)	Peer-reviewed; systematic approach; relevant to cultural/educational uses of AI	None
Raman et al.	2024	Fake news research trends, linkages to generative AI and SDGs	Bibliometric mapping	SDG 9 (Detection infrastructure), SDG 16 (Disinformation)	Peer-reviewed; explicit link between deepfakes, disinformation, and SDGs	Purely technical detection papers excluded
Romero-Moreno	2025	Deepfake detection in generative AI: A legal framework proposal	Legal analysis	SDG 16 (Governance, justice)	Peer-reviewed; legal focus on governance frameworks	Excluded commentaries without peer review
Hynek et al.	2025	Risks and benefits of deepfakes: Public attitudes in seven European countries	Empirical survey	SDG 16 (Trust, risk perception)	Large cross-national dataset; directly addresses public trust	None
Talebey Ahvanooney et al.	2025	DeepFakeR-MF: A risk framework for cyber (dis)trust behavior	Risk framework (model)	SDG 16 (Cyber trust)	Peer-reviewed; conceptual model of trust/distrust dynamics	Excluded papers focused only on algorithms
Furizal et al.	2025	Social, legal, and ethical implications of deepfake pornography	Systematic review	SDG 16 (Human rights, gendered harms)	Peer-reviewed; addresses social and legal harms	None
Daly, Hearn & Papageorgiou	2025	Sensemaking with AI: Trust in human-AI collaboration	Empirical (surveys + interviews)	SDG 9 (Creative partnerships), SDG 16 (Trust)	Empirical study with human participants; dual focus (innovation & trust)	Excluded studies lacking primary data
Atlam et al.	2025	SLM-DFS: A systematic literature map of deepfake spread on social media	Literature map	SDG 16 (Disinformation spread)	First large-scale map of 286 studies; directly relevant to disinformation	Excluded studies without societal/ethical scope
Creely & Blannin	2025	Creative partnerships with generative AI: Possibilities for education and beyond	Autoethnographic study	SDG 9 (Education, creative innovation)	Peer-reviewed; focuses on co-creation in education	None

Figure 4: Summary of reviewed literature and their SDG alignment.

Analytical Framework

The analysis employed a two-step coding strategy. First, each article was classified according to its contribution to SDG 9 or SDG 16. Second, findings were synthesized through thematic analysis, producing four overarching themes: (1) creative and industrial innovation, (2) disinformation and trust erosion, (3) governance and regulation, and (4) asymmetry of harms.

An interdisciplinary integration approach was applied, ensuring that technical, legal, and social perspectives were considered together rather than in isolation. This allowed deepfakes to be examined not only as algorithmic artifacts but also as socio-political phenomena.

Validity, Reliability, and Ethics

Validity was ensured by relying on peer-reviewed sources and triangulating findings across diverse methodologies (surveys, reviews, bibliometrics, legal studies). Reliability was supported through consistent coding categories linked directly to SDG 9 and SDG 16.

Although this research did not involve human participants, ethical considerations were central due to the sensitive themes of disinformation and non-consensual pornography. A victim-centered perspective was maintained, with language kept analytical and non-sensational.

4. Limitations

This methodology has limitations. The reliance on English-language studies excludes potentially valuable non-English literature. The 2021–2025 time frame focuses on recent debates, omitting earlier foundations. Empirical studies are concentrated in Europe and North America, leaving Global South perspectives underrepresented. Finally, as a qualitative synthesis, findings are interpretive rather than statistically generalizable.

5. Discussion

Deepfakes as Dual-Use Technology

The findings across the reviewed studies confirm that deepfakes are dual-use technologies. They are simultaneously tools of creativity and industrial modernization and vectors of exploitation and disinformation. This paradox is central to their significance in the context of the Sustainable Development Goals. Under SDG 9, deepfakes advance innovation in creative industries, education, and infrastructure. Under SDG 16, they corrode institutional trust, exploit vulnerable populations, and enable new forms of digital crime.

The paradox lies not in the technology itself but in the contexts of use, governance frameworks, and societal vulnerabilities. Evidence shows that AI supports artistic innovation when integrated as an augmentative tool [1]. In educational settings, AI can act as a co-creator, expanding the boundaries of human imagination [11]. Yet the same technology produces profound harm when used in non-consensual pornography [8]. Deepfakes also destabilize courts and journalism through the liar’s dividend, which enables authentic evidence to be dismissed as fabricated [5]. Together, these findings indicate that the promise or peril of deepfakes depends less on algorithms and more on regulation, consent, and ethical deployment [1], [5], [8], [11].

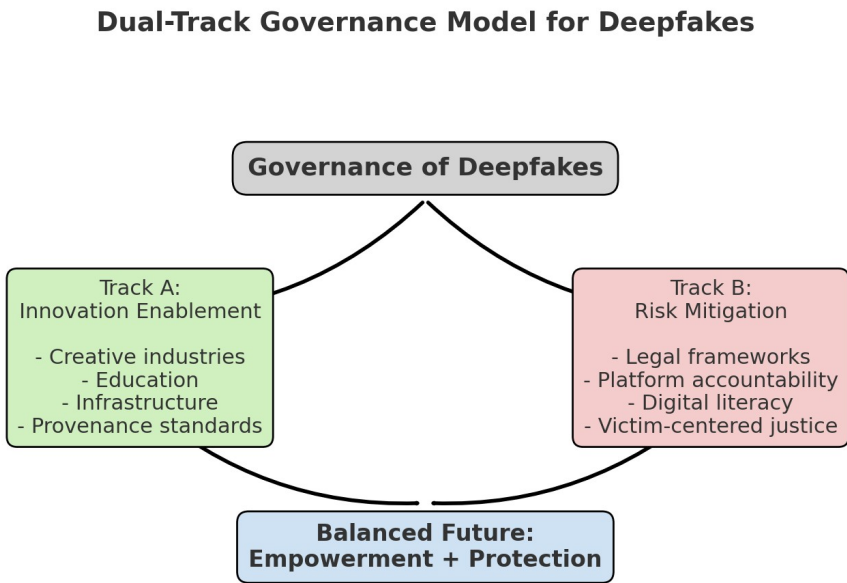


Figure 5: Show the dual-track governance strategy (Track A = Innovation Enablement, Track B = Risk Mitigation).Figure generated with the assistance of ChatGPT (OpenAI, 2025).

Governance Challenges and Policy Implications

Governance emerges as a recurring weakness in managing deepfakes. Evidence shows that while technical detection dominates academic work, governance, ethics, and regulatory interventions remain underexplored [10]. This imbalance leaves a gap between innovation and accountability. Comparative studies further demonstrate that approaches vary widely: the EU has taken a proactive regulatory stance, while the US remains fragmented, China enforces algorithmic labeling, and the UK focuses on platform accountability [5]. The result is a patchwork of regulatory strategies with limited harmonization [5], [10].

This fragmentation has consequences. Victims of deepfake pornography often lack effective legal recourse, since most jurisdictions do not recognize it as a distinct offense [8]. Publics also demand stronger protections, though attitudes vary across regions [6]. These findings point to a governance lag, where regulations fail to keep pace with innovation, leaving citizens exposed to abuse and manipulation [6], [8], [10].

The reviewed literature suggests that governance must evolve into a multi-layered model. Technical safeguards such as watermarking, provenance standards, and explainable AI provide the first line of defense. Legal frameworks, including criminalization of non-consensual deepfakes, international harmonization, and platform accountability, form the second. Finally, social interventions, such as digital literacy initiatives, awareness campaigns, and victim-centered justice mechanisms, ensure that protections reach vulnerable populations. Only by integrating these layers can governance balance the opportunities of SDG 9 with the justice commitments of SDG 16 [5], [6], [8], [10].

Ethical and Societal Considerations

Beyond governance, deepfakes raise ethical dilemmas about consent, authorship, and agency. Evidence from educational and creative contexts shows that AI can inspire collaboration, but only if authorship boundaries remain clear [9], [11]. In contrast, when deployed in exploitative contexts, deepfakes strip individuals of dignity and autonomy [8].

The contextual nature of ethics is evident: what counts as innovation in one domain can constitute abuse in another. Deepfakes used in film production or accessibility tools align with SDG 9, while those weaponized for harassment or disinformation undermine SDG 16. This reinforces the need for ethical frameworks that are dynamic, context-sensitive, and consent-based [8]–[11]. Without such frameworks, society risks normalizing synthetic media in ways that privilege efficiency and entertainment over human rights and justice.

Research Gaps

Despite rapid growth in the literature, several gaps remain:

1. Real-world testing of safeguards: Watermarking and provenance standards are widely proposed but rarely tested in live social media environments. Little is known about how users interpret and respond to these signals.
2. Consent verification: No standardized infrastructure exists for verifying digital consent in the creation and distribution of deepfakes. This gap weakens accountability and enforcement.
3. Institutional trust erosion: While theorized, causal links between deepfakes and declining trust in courts, elections, and journalism remain underexplored. Empirical longitudinal studies are needed.
4. Global South perspectives: Most research originates from Europe, North America, and China. Vulnerabilities in the Global South, where digital literacy and regulation may be weaker, are neglected.
5. Interdisciplinary collaboration: Studies remain siloed—computer scientists focus on detection, legal scholars on regulation, social scientists on attitudes. Few integrate these perspectives into comprehensive solutions.

Addressing these gaps will require coordinated, cross-disciplinary, and international research agendas.

6. Conclusion and future work

Deepfakes embody the paradox of artificial intelligence: they are simultaneously tools of creativity and engines of erosion. As demonstrated by the reviewed literature, they contribute to SDG 9 by enabling cultural innovation, educational enrichment, and infrastructural resilience. At the same time, they threaten SDG 16 by amplifying disinformation, undermining institutional trust, and facilitating image-based abuse.

The contributions of this article are threefold. First, it consolidates diverse debates into a coherent synthesis of deepfakes as dual-use technologies. Second, it frames findings explicitly through the lens of the SDGs, linking innovation and governance to global development priorities. Third, it advances the proposal of a dual-track governance model that enables responsible innovation while embedding safeguards against abuse.

The review also highlights critical gaps, including the absence of real-world testing of safeguards, the lack of standardized consent protocols, and the underrepresentation of Global South perspectives. These limitations underscore the need for more comprehensive and methodologically rigorous inquiry.

Future work will expand this analysis by conducting a full systematic review of deepfake research. Such a review will provide a broader evidence base, evaluate effectiveness of safeguards in practice, and incorporate global perspectives currently missing from the literature. This progression will ensure that governance frameworks evolve in tandem with innovation, determining whether deepfakes become instruments of empowerment or threats to peace and justice in the digital age.

Declaration on Generative AI

The authors confirm that the text, figures, and visualizations included in this article were partially generated with the assistance of OpenAI's ChatGPT (version GPT-5, 2025). The AI tool was used for language refinement, data visualization design (e.g., reproduction of figures based on bibliometric data from Raman et al., 2024), and structural editing. All intellectual contributions, conceptual design, data analysis, and interpretation of findings were performed by the authors. The AI-generated content was thoroughly reviewed and verified for accuracy, originality, and compliance with ethical research standards. The use of AI tools in this work aligns with the Declaration on Generative AI in Scholarly Writing and adheres to Elsevier's and IEEE's publication ethics guidelines.

References

- [1] N. Anantrasirichai and D. Bull, "Artificial intelligence in the creative industries: a review," *Artificial Intelligence Review*, vol. 54, pp. 427–481, 2021. doi: 10.1007/s10462-021-10039-7.
- [2] K. Feher and A. I. Katona, "Fifteen shadows of socio-cultural AI: A systematic review and future perspectives," *Futures*, vol. 132, p. 102817, 2021. doi: 10.1016/j.futures.2021.102817.
- [3] R. Raman, V. K. Nair, P. Nedungadi, A. K. Sahu, R. Kowalski, S. Ramanathan, and K. Achuthan, "Fake news research trends, linkages to generative artificial intelligence and sustainable development goals," *Heliyon*, vol. 10, no. 3, e24727, 2024. doi: 10.1016/j.heliyon.2024.e24727.
- [4] F. Romero-Moreno, "Deepfake detection in generative AI: A legal framework proposal to protect human rights," *Computer Law & Security Review*, vol. 58, p. 106162, 2025. doi: 10.1016/j.clsr.2025.106162.
- [5] N. Hynek, B. Gavurova, and M. Kubak, "Risks and benefits of artificial intelligence deepfakes: Systematic review and comparison of public attitudes in seven European countries," *Technology in Society*, vol. 74, p. 102611, 2025. doi: 10.1016/j.techsoc.2025.102611.
- [6] S. Taleby Ahvanooey, S. H. Islam, M. Abdur Razzaque, and J. W. Wong, "DeepFakeR-MF: A risk framework for cyber (dis)trust behavior in the age of synthetic media," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 107345, 2025. doi: 10.1016/j.engappai.2025.107345.

- [7] M. Furizal, M. D. K. Sari, M. A. Mahmud, and H. Fahmi, "Social, legal, and ethical implications of AI-generated deepfake pornography on digital platforms: A systematic literature review," *Digital Communications and Networks*, 2025 (in press). doi: 10.1016/j.dcan.2025.03.014.
- [8] S. J. Daly, G. Hearn, and K. Papageorgiou, "Sensemaking with AI: How trust influences human–AI collaboration in health and creative industries," *Social Sciences & Humanities Open*, vol. 11, 101346, 2025. doi: 10.1016/j.ssaho.2025.101346.
- [9] E.-S. Atlam, R. W. Atlam, A. A. Alzahrani, and A. E. Hassanien, "SLM-DFS: A systematic literature map of deepfake spread on social media," *Journal of King Saud University – Computer and Information Sciences*, 2025 (in press). doi: 10.1016/j.jksuci.2024.12.008.
- [10] E. Creely and J. Blannin, "Creative partnerships with generative AI. Possibilities for education and beyond," *Thinking Skills and Creativity*, vol. 49, 101643, 2025. doi: 10.1016/j.tsc.2025.101643.
- [11] W. Matli, "Extending the theory of information poverty to deepfake technology," *Int. J. Inf. Manage. Data Insights*, vol. 4, 100286, 2024, doi: 10.1016/j.jjimei.2024.100286.
- [12] M. Wazid, A.K. Mishra, N. Mohd, and A.K. Das, "A secure deepfake mitigation framework: Architecture, issues, challenges, and societal impact," *Cyber Security and Applications*, vol. 2, 100040, 2024, doi: 10.1016/j.csa.2024.100040.
- [13] J. Kaur, M. Singh, A. Sharma, and R. Gupta, "Deepfakes and the digital divide: Exploring the ethical and societal implications of synthetic media," *Comput. Stand. Interfaces*, vol. 90, 103909, 2024, doi: 10.1016/j.csi.2024.103909.