JOURNAL OF INFORMATION DISORDER AND MEDIA LITERACY(JIDM)

VOL.1-NO.1-2026

P ISSN: xxxx-xxx

E ISSN: xxxx-xxx

DEEPFAKES AND INFORMATION DISORDER: CHALLENGES FOR DIGITAL MEDIA LITERACY

Sana Afzal

BS, Department of Computer Science, The Cosmic Degree College (For Women) ,University of Okara

ABSTRACT

The rapid development of artificial intelligence has transformed digital media production through the creation of highly realistic synthetic images, videos, and audio recordings known as deepfakes. Although these technologies have legitimate applications in entertainment, education, and communication, they also create significant challenges for information credibility, public trust, and digital media literacy. Deepfakes can be used to impersonate individuals, manipulate public perceptions, spread disinformation, and undermine confidence in authentic digital evidence. This article examines the relationship between deepfakes and information disorder, focusing on detection challenges, media literacy, ethical concerns, and strategies for responsible digital communication. Previous research has demonstrated that deepfake generation and detection involve continuously evolving technical methods, making reliable identification increasingly difficult (Chesney & Citron, 2019; Mirsky & Lee, 2021). Research on synthetic media also highlights the importance of examining manipulated content within its social and informational context rather than relying exclusively on technical detection (Vaccari & Chadwick, 2020). This article adopts a qualitative, literature-based approach to analyze existing scholarship and identify major challenges for digital media literacy. The discussion emphasizes critical evaluation, source verification, visual and audio analysis, and awareness of AI-generated content. The article concludes that combining media literacy education, responsible AI development, transparent content labeling, and human verification can strengthen public resilience against deepfake-related information disorder.

KEYWORDS

Deepfakes, Information Disorder, Digital Media Literacy, Artificial Intelligence, Disinformation, Synthetic Media, Source Verification

INTRODUCTION

Digital media technologies have significantly changed how information is created, distributed, and consumed. Social media platforms, video-sharing applications, and online communication services allow users to publish and exchange information rapidly across geographical boundaries. However, these developments have also created opportunities for misleading content to circulate widely, particularly when manipulated images, videos, and audio recordings appear authentic. Deepfakes represent an important development in this environment because artificial intelligence can generate or modify media to imitate a person's appearance, voice, or actions. Chesney and Citron (2019) examined the legal and social implications of deepfakes, explaining how synthetic media can threaten privacy, reputation, democratic processes, and public trust. Their research provides an important foundation for understanding deepfakes as more than a technical problem: they also present challenges for information integrity and responsible communication.Deepfakes are commonly associated with deep learning techniques that generate or manipulate audiovisual content. These technologies can create realistic facial expressions, imitate voices, alter speech, or place individuals in situations that never occurred. Although some synthetic media is produced for entertainment, artistic expression, or accessibility, the same

capabilities can be misused to deceive audiences. Mirsky and Lee (2021) reviewed the development of deepfake creation and detection techniques, demonstrating that improvements in generation methods have encouraged corresponding efforts to identify manipulated content. This continuing technological competition creates difficulties for journalists, educators, researchers, and ordinary users who need to assess the authenticity of digital media. Consequently, digital media literacy must address both the capabilities of synthetic technologies and the limitations of methods used to detect them.

Information disorder provides a broader framework for understanding the consequences of deepfakes. Misleading digital content does not necessarily depend on completely fabricated material; authentic photographs, recordings, and statements can also be presented in deceptive contexts. Wardle and Derakhshan (2017) distinguished between misinformation, disinformation, and malinformation, providing a framework for examining how information can be misleading, deliberately deceptive, or harmful through its use and distribution. Deepfakes can contribute to disinformation when synthetic content is deliberately created or circulated to mislead audiences. They can also contribute to wider information disorder when their existence encourages people to question authentic evidence. This means that assessing deepfakes requires attention to the content itself, the context in which it appears, the source responsible for its distribution, and the intentions or consequences associated with its use.

The increasing availability of synthetic media has made it more difficult for audiences to rely on visual and auditory evidence alone. Previously, a video or voice recording might have been regarded as strong evidence that an event occurred or that a person made a particular statement. However, convincing synthetic content weakens this assumption. Vaccari and Chadwick (2020) examined the implications of political deepfakes for misinformation, uncertainty, and trust in news on social media. Their research highlights how manipulated media can create uncertainty even when audiences cannot confidently determine whether a particular item is authentic. This uncertainty presents an important challenge for digital media literacy because individuals must learn not only how to question suspicious content but also how to avoid automatically dismissing genuine evidence as fake.Digital media literacy is therefore essential for helping users respond critically to synthetic content. Media literacy involves evaluating sources, examining evidence, understanding how digital content is produced, and recognizing the limitations of online information. In the context of deepfakes, these skills include checking the original source of a video, comparing claims with reliable reporting, examining the date and context of a recording, and consulting independent verification services. Technical indicators such as unusual facial movements or inconsistent audio may sometimes be useful, but they should not be treated as definitive proof because synthetic media continues to improve. Mirsky and Lee (2021) emphasize the continuing challenges associated with deepfake detection, supporting the need for approaches that combine technical analysis with contextual investigation and human judgment.

Another challenge concerns the relationship between deepfakes, public trust, and social media distribution. Synthetic content can spread quickly when it attracts attention, generates emotional reactions, or supports existing political and social beliefs. Users may share misleading videos before their authenticity has been established, allowing false impressions to develop even when

corrections are subsequently published. The problem can be particularly serious when synthetic media targets public figures, vulnerable individuals, or sensitive social issues. Chesney and Citron (2019) discuss the potential harms associated with malicious deepfakes, while Vaccari and Chadwick (2020) demonstrate why the uncertainty created by manipulated media deserves attention in discussions of public trust. These concerns highlight the importance of responsible sharing practices, clear corrections, and public education about synthetic content.The challenges of deepfakes also extend beyond individual users to educational institutions, journalists, technology companies, and policymakers. Educators need suitable resources to teach students how to evaluate audiovisual evidence, while journalists must verify media before publishing or reproducing it. Technology companies face the challenge of identifying manipulated content at scale without incorrectly labeling authentic material. Policymakers must consider how to address harmful uses of synthetic media while protecting legitimate expression and creative applications. Technical detection systems, provenance information, and content labeling can support these efforts, but no single approach can address every form of manipulation. Effective responses therefore require cooperation among technological developers, educational institutions, media organizations, and the public.This article examines deepfakes and information disorder through the perspective of digital media literacy. It focuses on the risks associated with synthetic media, the limitations of detection techniques, the importance of source verification, and the role of education in promoting responsible information practices. Using a qualitative review of existing scholarship, the article seeks to explain how individuals and institutions can respond to deepfake-related challenges without treating all digital content as unreliable. Its central argument is that digital media literacy should combine critical thinking, contextual verification, awareness of AI-generated content, and responsible communication. Such an approach can help strengthen information credibility while enabling users to engage with digital technologies more confidently and thoughtfully.

RESEARCH OBJECTIVES

1. To examine how deepfakes contribute to information disorder in digital media environments.
2. To identify the technological, social, and educational challenges associated with detecting deepfakes.
3. To explore the role of digital media literacy in promoting critical evaluation and responsible engagement with synthetic media.

RESEARCH QUESTIONS

1. How do deepfakes contribute to information disorder and affect trust in digital media?
2. What technological, social, and educational challenges limit the identification of deepfakes?
3. How can digital media literacy help individuals evaluate and respond to deepfake content responsibly?

LITERATURE REVIEW

Research on deepfakes has developed across artificial intelligence, communication studies, digital media, and information ethics. Chesney and Citron (2019) examined the potential consequences of malicious synthetic media for privacy, reputation, political communication, and public trust. Their analysis demonstrates that deepfakes can create serious problems when manipulated content is used to deceive audiences or damage individuals' reputations. Wardle and Derakhshan (2017) provide a broader framework for understanding information disorder by distinguishing misinformation, disinformation, and malinformation. These concepts help explain why deepfakes should be studied not only as manipulated audiovisual objects but also as elements of wider information environments in which content is produced, circulated, interpreted, and used to influence public opinion.Technical research has focused on the development of deepfake generation methods and the challenges of detecting manipulated content. Mirsky and Lee (2021) reviewed deepfake creation and detection techniques, demonstrating how improvements in synthetic media generation have encouraged the development of increasingly sophisticated detection approaches. Abbas and Taeihagh (2024) conducted a systematic review of deepfake generation and detection techniques, examining approaches involving images, audio, and video, as well as the policy challenges associated with their use. Their research indicates that technical detection can contribute to identifying manipulated media, but the effectiveness of individual approaches depends on the type of content, the available data, and the conditions under which a model is evaluated. This literature supports the need to combine technical detection with contextual verification and public awareness.Human ability to recognize deepfakes is another important area of investigation. Diel et al. (2024) conducted a systematic review and meta-analysis of 56 papers examining human performance in deepfake detection. Their findings indicate that human detection performance can be limited, with accuracy often approaching chance levels across different types of synthetic media. The authors also found evidence that strategies intended to improve detection can increase performance. These findings challenge the assumption that viewers can reliably identify manipulated content simply by watching a video or listening to a recording. For digital media literacy, the implication is that users should be taught structured verification practices rather than relying exclusively on visual impressions, intuition, or confidence in their own judgments.

Recent research has also examined how people perceive and respond to deepfakes in everyday digital environments. Wang (2026), in a systematic review of empirical deepfake research, mapped findings from 261 quantitative and qualitative studies described in 219 articles. The review identified a need for more geographically diverse research, particularly in regions that have received comparatively limited attention. This finding is relevant to digital media literacy because public understanding of synthetic content may differ according to language, cultural context, educational opportunities, and digital media practices. Research that concentrates on a limited number of countries may not fully explain how people in other settings recognize or interpret manipulated information. Therefore, media literacy strategies should be evaluated across diverse communities rather than assuming that a single educational model will be equally effective everywhere.Research specifically examining young people's encounters with synthetic media provides further insight into these challenges. Everyday Encounters with Deepfakes: Young People's Media and Information Literacy Practices with AI-Generated Media (2025) examined interviews with 20 participants aged 14–15. The study explored young people's

experiences with AI-generated media, their understanding of deepfakes, and their responses to such content in everyday life. Its findings highlight the importance of understanding how young people encounter and interpret synthetic content, rather than assuming that their familiarity with digital platforms automatically provides the ability to recognize manipulated media. This research supports educational approaches that connect media literacy to young people's actual digital experiences and encourage them to investigate the origins, purposes, and credibility of online content.

The relationship between deepfakes and information disorder also involves the wider social consequences of uncertainty. Vaccari and Chadwick (2020) examined political deepfakes and their implications for misinformation, uncertainty, and trust in news on social media. Their research demonstrates why synthetic media can be damaging even when audiences do not necessarily believe every manipulated item. The existence of convincing fake content can create doubt about authentic recordings and weaken confidence in reliable information. This phenomenon highlights the importance of source verification, independent corroboration, and transparent correction practices. Media literacy should therefore teach users both to recognize potential manipulation and to avoid dismissing genuine evidence without sufficient grounds.

Recent systematic reviews further emphasize the need to integrate technical detection, educational awareness, and responsible governance. Khan et al. (2025) surveyed multimedia-enabled deepfake detection methods and examined developments in machine learning, deep learning, and multimodal systems. Their review discussed technical performance, robustness, and emerging approaches intended to improve detection. Mandal et al. (2026) also examined deepfakes and AI-generated disinformation, considering detection techniques, datasets, ethical risks, and policy responses. Together, this research indicates that technical systems can support the identification of synthetic media, but challenges involving evolving generation techniques, limited dataset diversity, and real-world application remain important. Digital media literacy can complement these systems by helping users interpret warnings, verify sources, and understand the limits of automated detection.

Overall, the literature demonstrates that deepfakes create interconnected technological, social, and educational challenges. Technical studies emphasize the importance of improved detection systems, while communication research highlights uncertainty, trust, and the circulation of misleading content. Research on human performance and young people's media practices further suggests that users require structured guidance to evaluate synthetic media effectively. The central implication is that digital media literacy should combine source verification, contextual analysis, awareness of AI-generated content, and responsible information sharing. Further research is needed to assess how these strategies perform across different age groups, languages, cultural contexts, and digital environments.

METHODOLOGY

This article adopts a qualitative, literature-based methodology to examine deepfakes and their implications for information disorder and digital media literacy. Relevant scholarly publications were identified through Google Scholar and academic publishing databases using search terms

such as deepfakes, synthetic media, information disorder, deepfake detection, and digital media literacy. Priority was given to peer-reviewed research, systematic reviews, and studies published between 2017 and 2026 that directly addressed the article's research questions. The selected literature was examined to identify recurring themes concerning synthetic media generation, detection limitations, public trust, source verification, and media literacy education. Findings from technical studies were considered alongside research on human perceptions and communication practices. The study relies on existing published research and does not involve original interviews, surveys, or experiments. Accordingly, its conclusions represent a qualitative synthesis of available literature rather than new empirical measurements.

RESEARCH DESIGN

The study employs a qualitative descriptive research design based on the review and thematic comparison of existing academic literature. This design is appropriate for examining the interconnected technological, social, and educational dimensions of deepfake-related information disorder. The literature is organized around three principal areas: the contribution of deepfakes to misinformation and declining information trust, the limitations of human and automated detection, and the role of digital media literacy in evaluating synthetic content. Selected publications are compared according to their research objectives, methodologies, principal findings, and relevance to the research questions. Particular attention is given to differences between technical detection research and studies examining human responses to manipulated media. This design enables the identification of common findings, unresolved challenges, and opportunities for further research without claiming to establish new causal relationships.

DATA ANALYSIS

The study uses qualitative thematic analysis to examine the selected literature. Relevant findings are grouped into themes concerning deepfake generation, detection accuracy, information credibility, public trust, digital verification, and media literacy education. The analysis compares how technical studies describe the strengths and limitations of detection systems and how communication studies explain the social consequences of synthetic content. Research on human detection is examined to identify difficulties associated with relying on visual impressions or intuition, while studies of media literacy are considered for their implications for source verification and responsible information sharing. The identified themes are interpreted in relation to the three research questions. Because the article does not collect primary data, it does not calculate original detection accuracy, conduct statistical tests, or claim experimental validation of educational strategies. Instead, it develops a structured interpretation of existing findings and identifies areas where further empirical investigation is needed.

FINDINGS

The literature review indicates that deepfakes have introduced significant challenges for information credibility because they can make fabricated audiovisual content appear authentic. Synthetic videos, images, and audio recordings can be used to impersonate individuals, misrepresent events, and circulate misleading claims. Chesney and Citron (2019) highlight the

potential consequences for privacy, reputation, and public trust, while Wardle and Derakhshan (2017) provide a broader framework for understanding the role of manipulated content in information disorder. These studies suggest that deepfakes should be examined not only as technological products but also in relation to the contexts in which they are created, distributed, and interpreted.A second finding concerns the limitations of human detection. Diel et al. (2024) found that people frequently struggle to distinguish genuine media from deepfakes reliably. This indicates that audiences cannot always identify manipulated content through visual inspection or listening alone. Although some inconsistencies may provide clues, the absence of obvious errors does not establish authenticity. Consequently, digital media literacy should encourage users to investigate the original source, compare independent reports, verify dates and context, and seek credible corroboration before sharing potentially consequential material.

The review also identifies continuing technical challenges in automated detection. Abbas and Taeihagh (2024) examined detection techniques across different forms of synthetic media and identified challenges involving detection methods and their practical implementation. As generation techniques evolve, systems trained on particular datasets may not perform equally well on unfamiliar forms of manipulation. Differences in media quality, language, and content type can further complicate identification. These limitations indicate that detection tools should be evaluated under realistic conditions and that their results should be interpreted cautiously rather than treated as definitive judgments.Another finding concerns the relationship between deepfakes and public trust. Vaccari and Chadwick (2020) demonstrated the importance of uncertainty surrounding political deepfakes and its implications for trust in social media news. Manipulated content can mislead audiences directly, but it can also encourage people to doubt authentic recordings and reliable reporting. This creates a broader information problem in which uncertainty itself can weaken public confidence. Effective responses therefore require transparent verification, timely corrections, and clear communication about the evidence supporting a claim.The literature further indicates that digital media literacy is essential for responding to deepfake-related risks. Young people and adults need to understand how synthetic content is created, why it may be misleading, and how its authenticity can be investigated. Research on young people's everyday encounters with AI-generated media highlights the importance of examining their actual experiences and understanding of deepfakes rather than assuming that digital familiarity guarantees media literacy (Everyday Encounters with Deepfakes, 2025). Educational activities should therefore include practical source checks, discussion of manipulated media examples, and guidance on responsible information sharing. These activities can help users develop more reflective habits when encountering unfamiliar audiovisual content.

Finally, the findings suggest that no single intervention can address every dimension of deepfake-related information disorder. Technical detection systems can support verification, media literacy can improve users' understanding, and responsible platform practices can help limit the spread of deceptive content. However, these approaches need to operate together. Educators, journalists, researchers, technology developers, and policymakers each have a role in promoting reliable information practices. The review therefore supports a combined approach

that integrates technical innovation, human judgment, media literacy, and transparent communication.

CONCLUSION

Deepfakes represent a significant challenge for digital media literacy because they can blur the distinction between authentic evidence and artificially generated content. Their potential misuse extends beyond fabricated videos and audio recordings to broader concerns involving misinformation, personal reputation, public trust, and the credibility of digital communication. The reviewed literature demonstrates that both human judgment and automated detection have limitations, making it necessary to adopt a comprehensive approach to evaluating synthetic media. Users should not rely exclusively on visual impressions or automated classifications when deciding whether a recording is genuine.The article also highlights the importance of teaching practical verification skills. Digital media literacy should help individuals examine original sources, compare independent information, evaluate contextual evidence, and recognize the limitations of detection technologies. Educational institutions can support these skills through structured learning activities, while journalists and digital platforms can contribute through transparent verification practices and responsible handling of manipulated content. Such measures can help reduce the risks associated with misleading synthetic media without encouraging users to distrust all digital evidence.In conclusion, addressing deepfakes and information disorder requires cooperation among educators, media organizations, technology developers, policymakers, and the public. Future research should investigate the long-term effectiveness of media literacy interventions, the performance of detection tools across different languages and cultural contexts, and the social effects of synthetic content. By combining critical thinking, reliable verification, technical innovation, and responsible digital participation, society can strengthen information integrity and respond more effectively to the challenges posed by deepfakes.

REFERENCES

Abbas, F., & Taeihagh, A. (2024). Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. Expert Systems with Applications, 252, Article 124260.

Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107(6), 1753–1820.

Diel, A., Lalgi, T., Schröter, I. C., MacDorman, K. F., Teufel, M., & Bäuerle, A. (2024). Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. Computers in Human Behavior Reports, 16, Article 100538.

Mandal, U., Setua, S. K., & Sarma, S. S. (2026). Deepfakes and beyond in the era of AI-generated disinformation: A systematic literature review of emerging technologies, challenges, and policy recommendations. Journal of Visual Communication and Image Representation, 117, Article 104748.

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. ACM Computing Surveys, 54(1), Article 7.

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).

Wang, S. (2026). Mapping the landscape of deepfake research: A systematic review of empirical studies. Big Data & Society.

Wardle, C., & Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policymaking. Council of Europe.