Misinformation to Mitigation: Strategies for Combating Deepfakes on Social Media

1st Marco Trisna Omar Farrasy[1], 2nd Moch Fuad Nasvian[2]

{marcotof17@gmail.com[1], nasvian@umm.ac.id[2]}

Universitas Muhammadiyah Malang, +62877-6182-8094[12]

Abstract. Social media has become a primary news source, but deepfakes—synthetic media generated by advanced AI pose significant threats, including political manipulation, non-consensual imagery, misinformation, and financial fraud. These applications erode trust, fuel misinformation, and deepen societal polarization. This study systematically reviewed literature published between 2018 and 2023 using PRISMA guidelines to evaluate these threats and propose the 4C Model of Deepfake Mitigation: Context (understanding socio-digital environments), Content (ensuring media authenticity), Community (promoting literacy and collaboration), and Control (enforcing regulations and platform accountability). Emphasizing interdisciplinary approaches, the research highlights media literacy, user empowerment, and robust platform policies as key solutions. Insights are provided for policymakers, technologists, and educators to address deepfake challenges and protect digital ecosystems.

Keywords: Deepfake, Social Media, Artificial Intelligence, Systematic Literature Review, PRISMA

1 Introduction

This is preliminary research about Combating Deepfakes on Social Media. By this paper, researchers propose the concept of deepfakes mitigation strategies on social media. Social media platforms have evolved significantly over the past decade. It transformed from simple tools for personal communication to becoming dominant sources of news and information for users worldwide [1], [2], [3], [4]. This shift has enhanced global connectivity but also facilitated the rapid spread of misinformation [5], [6]. Among the most advanced and troubling manifestations of misinformation today are the use of artificial intelligence (AI) to generate realistic synthetic media called “Deepfakes”. These highly convincing digital fabrications have redefined the landscape of fake news, raising complex ethical, social, and technological challenges [7], [8], [9].

Deepfakes are used in diverse domains, ranging from political manipulation to non-consensual imagery, misinformation campaigns, and financial fraud [9], [10], [11], [12]. By creating hyper-

ICHSS 2024, December 11, Surakarta, Indonesia

DOI 10.4108/eai.11-12-2024.2360426

realistic yet entirely false representations of individuals and events, deepfakes erode societal trust, amplify misinformation, and contribute to social polarization [13], [14], [15]. On social media, where information spreads rapidly and widely, the disruptive potential of deepfakes is magnified, posing significant risks to individual privacy, democratic processes, and societal cohesion [16], [17].

Addressing these challenges requires a thorough understanding of the threats posed by deepfakes as well as the strategies available to mitigate their impact. Existing research highlights the need for multi-disciplinary approaches combining technological innovation, media literacy, and regulatory frameworks to combat this growing issue effectively. Therefore, to contribute to this body of knowledge by conducting a systematic literature review of research on deepfakes is crucial to analyze research data over the past few years.

Therefore, this research will address two key research questions:

1. How does existing research describe the threat posed by deepfakes on social media?
2. What strategies are proposed to mitigate the social impact of deepfakes?

By addressing these questions, this study will provide a comprehensive understanding of the multifaceted impact of deepfakes on social media and offer actionable insights to users.

2 Research Method

This research conducted a literature review search and analysis from Scopus database using Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, which is a widely used framework for synthesizing literature review [28]. Here are the steps that we followed.

3.1 Data Source and Literature Search

This research conducted an initial search of the literature through the Scopus database. The time frame for the reviewed literature is from 2018 to 2023. The combinations of keywords *deepfakes and social media* were used to identify the papers in Scopus database. The more detailed search syntax used for this research is shown in Table 1. Papers included in this review are English journals only and focusing on Social Studies. For the methods there were no restrictions in quantitative, qualitative, or mixed methods.

Table 1. Search Strategy Syntax.

Database	Syntax	Number of Articles
Scopus	(TITLE-ABS-KEY("Deep fakes" AND "Social media") AND (LIMIT-TO (SRCTYPE,"j"))	10

	AND (LIMIT-TO (LANGUAGE,"English")) AND (LIMIT-TO (SUBJAREA,"SOCI")))	

3.2 Study Selection

The article search based on the inclusion/exclusion criteria identified a total of 10 articles (Figure 1). The titles and abstracts of the articles were reviewed if they met the criteria focusing on the use of AI with deepfakes on social media. The purpose of the review is to help contextualize the key features of Deepfakes on Social Media. After applying these criteria, 5 articles remained for full-text review.

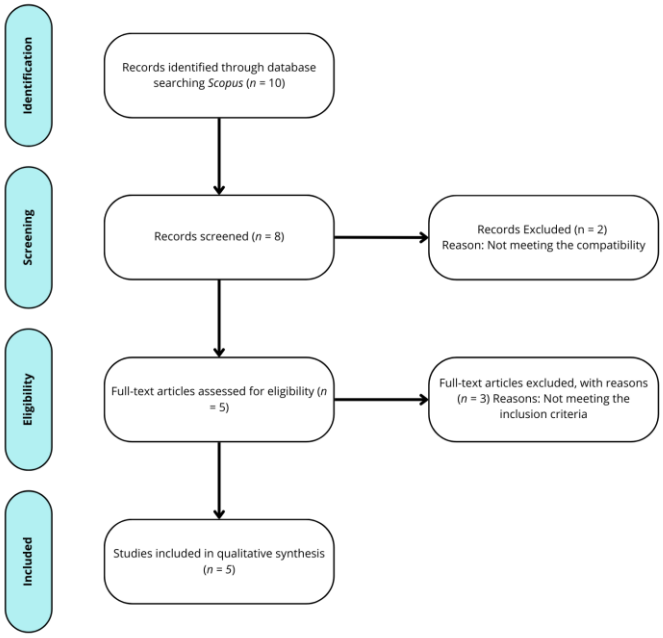

Fig. 1. Study Selection Diagram.

3.3 Approach to Analysis and Synthesis

The review analyzed 10 full-text articles on deepfake use in social media, excluding non-English or irrelevant studies. After consensus, 5 articles were synthesized qualitatively. Using Thomas & Harden (2008) thematic synthesis framework, the process involved three stages: line-by-line coding via Google Sheets, developing descriptive themes, and

generating analytical themes. Themes and coding were collaboratively reviewed by both authors, with findings detailed in the next section.

3 Result and Discussion

This section presents the results of thematic analysis of the literature on Deepfakes on Social Media to answer the research questions.

Table 2. Threats and Mitigation Strategies of the Included Studies.

Authors	Year	The Threat Posed by Deepfakes on Social Media			Strategies to Mitigate Social Impact of Deepfakes		
		Examples and Areas of Threat	Impact on Trust and Society	Impact on Misinformation	Educational Interventions and Strategies	Policy and Regulation	User Behavior and Awareness
Ahmed, Saifuddin	2022	Political Manipulation	-	-	-	-	Understanding Motivations Social Media Habits
Stover, Dawn	2018	Political Threat	Erosion of Trust Societal Impact	Proliferation of Fake News Mitigation Efforts	Media Literacy Training Parental Guidance	Comprehensive Approach Transparency and Accountability Incentive Structures	Engagement with Reliable Sources Incentives for Verification

							Community-Based Solutions
Nightingale, Sophie J; Wade, Kimberley A.	2022	Non-consensual Imagery Misinformation Campaigns Financial Fraud and Identity Theft	-	Amplification of False Information	Raising Awareness Critical Thinking Skills Media Literacy	Accountability for Social Media Companies Regulatory Frameworks	Understanding Motivated Reasoning Promoting Analytical Engagement Awareness of Manipulation Techniques
Ahmed, Saifuddin	2023	-	-	-	Digital Literacy Interventions Mini-Games for Discerning Deepfakes	Government Restrictions on Political Deepfakes Electoral Guidelines for	Cognitive Ability and Skepticism Higher Cognitive Individuals' Restraint

					Spot the Misinformation Games	Political Campaigns	Social Media News Skepticism
El Mokadem, Sarah Shawky	2023	-	Erosion of Trust	-	Media Literacy Programs Inoculation Theory Application Longitudinal Studies	Digital Platform Regulation Collaboration with Social Media Platforms	-

3.1 The Threat Posed by Deepfakes on Social Media

Deepfakes pose threats across political, social, and economic domains. In politics, they manipulate public opinion and disrupt democratic processes, as seen in a doctored video of Nancy Pelosi and real-time manipulations of George W. Bush, raising concerns about platform accountability and diplomatic stability [10], [11]. Non-consensual sexual imagery, targeting mostly women, constitutes another misuse, with 96% of online deepfake videos in 2019 being pornographic, violating privacy and dignity [20]. Deepfakes also fuel misinformation campaigns by spreading fabricated content using accessible tools like FakeApp, eroding trust in credible sources [20]. Economically, they facilitate fraud and identity theft through manipulated media [20].

Deepfakes undermine trust in reliable information, threatening democracy and national security, while intensifying societal polarization and weakening cohesion [11], [25]. The spread of misinformation is amplified by accessible tools like GANs and FakeApp, which create and disseminate convincing false content, making misinformation harder to counter [11]. Mitigating these challenges requires authentication technologies and interdisciplinary collaboration to ensure media integrity and combat the rapid spread of fake content [20].

4.2 Strategies to Mitigate Social Impact of Deepfakes

4.2.1 Educational Interventions and Strategies

Media literacy programs empower individuals to evaluate media critically, fostering informed decision-making and reducing susceptibility to manipulation, especially among the youth [11]. These programs, when integrated into school curricula, enhance critical thinking, enabling users to assess content authenticity by examining its source and purpose while addressing cognitive biases like partisan reasoning [20], [25]. Interactive tools such as "Bad News" and public awareness campaigns further boost resilience against misinformation by teaching users to identify manipulation techniques [22]. Parents and community-specific strategies play crucial roles in fostering early media literacy and vigilance [11], [25]. Inoculation Theory, which builds resilience by exposing individuals to weakened misinformation, complements these strategies [25].

4.2.2 Policy and Regulation

Tackling deepfakes demands collaboration between technologists, policymakers, and social scientists to design solutions that rebuild trust in digital systems [11]. Regulatory frameworks should prioritize transparency by labeling manipulated content and standardizing identification practices. Platforms must adopt safeguards like AI moderation tools and user-reporting mechanisms to prevent misuse [20], [25]. Governments should enforce ethical guidelines to restrict deepfake use in political

messaging, ensuring democratic integrity [11], [22]. Additionally, partnerships between platforms and governments are essential for tracking misinformation patterns and implementing countermeasures [25].

4.2.3 User Behavior and Awareness

Users often share deepfakes due to political agendas, entertainment, or social pressures, with FOMO being a key motivator [10]. Promoting engagement with credible sources and incentivizing content verification can reduce misinformation spread. Community-driven solutions, such as peer-mediation, encourage responsible sharing and enhance digital literacy [11]. Raising awareness about deepfake creation and manipulation helps users recognize red flags, while teaching motivated reasoning reduces bias in content perception [20]. Cognitive abilities also influence user responses, with lower cognitive ability linked to skepticism and higher ability fostering critical evaluation [22]. Interactive tools like "Bad News" remain effective in empowering users to detect misinformation and foster an informed digital community [22].

Discussion

This research initially set out to explore how deepfakes are perceived as threats on social media and to analyze existing mitigation strategies, intending to provide a comprehensive overview of the issue. The focus was on identifying how various academic and practical discussions addressed the societal impact of deepfakes. However, the analysis revealed that while terms like "threats" and "mitigation strategies" were frequently referenced, they were often fragmented or lacked a cohesive framework that tied the discussion together in a structured way. This gap in the literature led to the development of the 4C Model of Deepfake Mitigation.

The 4C Model of Deepfake Mitigation offers a practical and well-rounded approach to tackling the growing threats of deepfakes. It revolves around four key areas: Context, Content, Community, and Control, each addressing different aspects of the challenge.

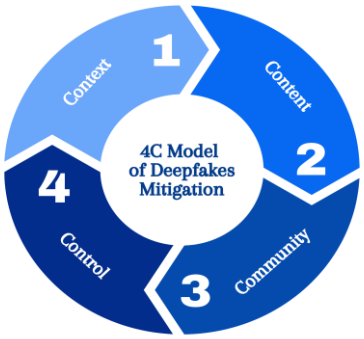

Fig. 2. 4C Model of Deepfakes Mitigation.

3.1 Context: Understanding Socio-Digital Environments

This pillar focuses on analyzing the ecosystems where deepfakes proliferate, characterized by low media literacy, high social media usage, and inadequate regulation. Contextual factors like societal divisions and cultural differences influence the spread of deepfakes. Tailored media literacy programs can empower audiences to recognize and resist manipulation.

3.2 Content: Ensuring Media Authenticity

The Content pillar addresses the erosion of trust caused by deepfakes through fake news and non-consensual imagery. Solutions include technologies like blockchain verification and digital watermarking to validate content. Mandatory labeling of AI-generated media by platforms and political campaigns is also essential to rebuild public confidence in digital information.

3.3 Community: Empowering Users

Community emphasizes equipping individuals and groups to combat deepfakes. Tools like the "Bad News" game, teaching misinformation detection, should be integrated into education and public awareness efforts. Community-driven moderation systems can enable users to report and mitigate harmful content, fostering safer online spaces.

3.4 Control: Regulating Platforms and Collaboration

Control involves holding digital platforms accountable for addressing deepfake content. Regulatory frameworks should enforce penalties for non-compliance and encourage platforms to adopt adaptive solutions. Collaboration among technologists, policymakers, and researchers is crucial for developing robust and scalable responses to deepfakes.

4 Conclusion

This study aimed to explore the threats posed by deepfakes on social media and identify effective strategies for mitigating their societal impact, as outlined in the research questions. However, while much of the existing literature discusses individual aspects of deepfake threats and solutions, it lacks a unified framework. This gap in research led to the development of the 4C Model of Deepfake Mitigation—a structured approach addressing four key areas: Context, Content, Community, and Control.

The findings of this study highlight the multifaceted risks of deepfakes, from political manipulation and non-consensual imagery to misinformation campaigns and financial fraud.

These challenges erode trust, deepen societal polarization, and undermine democratic processes. The proposed 4C Model integrates media literacy, technological innovation, community-driven action, and policy reform into a cohesive framework for addressing these issues. It emphasizes understanding the socio-digital environment, ensuring content authenticity, empowering users, and holding platforms accountable.

This research underscores the critical need for collaboration among policymakers, educators, technologists, and social media platforms to implement adaptive strategies and ethical governance. By fostering media literacy, advancing authentication technologies, and promoting transparent regulatory measures, stakeholders can enhance societal resilience against deepfakes and restore trust in digital ecosystems.

Future research should build on this study by testing the 4C Model in diverse cultural and technological contexts, refining its elements to address emerging challenges in synthetic media. As deepfake technologies continue to evolve, a proactive and collaborative approach will be essential to safeguarding the integrity of information and the well-being of digital communities.

References

[1] S. A. Afaq *et al.*, "Social Media Revolution," 2024, pp. 257–272. doi: 10.4018/979-8-3693-9235-5.ch013.

[2] A. Saima, N. Iqbal, and R. Ishaq, "Social Media as a News Source: An Analysis of Facebook," *Global Multimedia Review*, vol. V, no. I, pp. 24–46, Dec. 2022, doi: 10.31703/gmmr.2022(V-I).03.

[3] X. Li, H. Pan, and J. Yao, "Analyzing the Transformation of Journalism Practices Driven by the Rise of Social Media Platforms," *MEDAAD*, vol. 2023, pp. 18–25, Mar. 2023, doi: 10.70470/MEDAAD/2023/003.

[4] A. Moallem, "Trust in News and Information in Social Media," 2020, pp. 129–134. doi: 10.1007/978-3-030-52581-1_17.

[5] M. Alsaid, S. Parvathi Panguluri, and S. Hawamdeh, "Combating Misinformation on Social Media Using Social Noise and Social Entropy as a Measure of Uncertainty," *Proceedings of the Association for Information Science and Technology*, vol. 61, no. 1, pp. 25–35, Oct. 2024, doi: 10.1002/pra2.1005.

[6] S. E. V. S. Pillai, "Analyzing Network Characteristics for Misinformation Detection in Online Social Media," in *2024 4th International Conference on Data Engineering and Communication Systems (ICDECS)*, IEEE, Mar. 2024, pp. 1–6. doi: 10.1109/ICDECS59733.2023.10503325.

[7] I. Amerini *et al.*, "Deepfake Media Forensics: State of the Art and Challenges Ahead," Aug. 2024, [Online]. Available: http://arxiv.org/abs/2408.00388

[8] Vishal Gawali, Chaturdhan Chaubey, Mahesh Gaikwad, Akash Gidde, and Nilesh Bhelkar, "Study on AI Generated Fake-Media Detection," *International Research Journal on Advanced Engineering and Management (IRJAEM)*, vol. 2, no. 10, pp. 3181–3185, Oct. 2024, doi: 10.47392/IRJAEM.2024.0470.

[9] M. Momeni, "Artificial Intelligence and Political Deepfakes: Shaping Citizen Perceptions Through Misinformation," *Journal of Creative Communications*, Oct. 2024, doi: 10.1177/09732586241277335.

[10] S. Ahmed, "Disinformation Sharing Thrives with Fear of Missing Out among Low Cognitive News Users: A Cross-national Examination of Intentional Sharing of Deep

Fakes," *J Broadcast Electron Media*, vol. 66, no. 1, pp. 89–109, Jan. 2022, doi: 10.1080/08838151.2022.2034826.

[11] D. Stover, "Garlin Gilchrist: Fighting fake news and the information apocalypse," *Bulletin of the Atomic Scientists*, vol. 74, no. 4, pp. 283–288, Jul. 2018, doi: 10.1080/00963402.2018.1486618.

[12] C. P. Walker, D. S. Schiff, and K. J. Schiff, "Merging AI Incidents Research with Political Misinformation Research: Introducing the Political Deepfakes Incidents Database," 2024, [Online]. Available: www.aaai.org

[13] D. Cavedon-Taylor, "Deepfakes: a survey and introduction to the topical collection," *Synthese*, vol. 204, no. 1, p. 14, Jun. 2024, doi: 10.1007/s11229-024-04634-8.

[14] J. Twomey, D. Ching, M. P. Aylett, M. Quayle, C. Linehan, and G. Murphy, "Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine," *PLoS One*, vol. 18, no. 10, p. e0291668, Oct. 2023, doi: 10.1371/journal.pone.0291668.

[15] S. Alanazi, S. Asif, and I. Moulitsas, "Examining the Societal Impact and Legislative Requirements of Deepfake Technology: A Comprehensive Study," *International Journal of Social Science and Humanity*, 2024, doi: 10.18178/ijssh.2024.14.2.1194.

[16] A. V. Nadimpalli and A. Rattani, "Social Media Authentication and Combating Deepfakes using Semi-fragile Invisible Image Watermarking," *Digital Threats: Research and Practice*, Oct. 2024, doi: 10.1145/3700146.

[17] Vandana and K. Chaturvedi, "Illusion or Reality: Analyzing Sentiments on Deepfakes," in *2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, IEEE, Aug. 2024, pp. 1207–1210. doi: 10.1109/ICESC60852.2024.10689970.

[18] B. Chu, W. You, Z. Yang, L. Zhou, and R. Wang, "Protecting World Leader Using Facial Speaking Pattern Against Deepfakes," *IEEE Signal Process Lett*, vol. 29, pp. 2078–2082, 2022, doi: 10.1109/LSP.2022.3205562.

[19] N. Mrvić-Petrović, "Criminal law approach to regulating non-consensual pornographic deepfake," *Bezbednost, Beograd*, vol. 66, no. 2, pp. 5–23, 2024, doi: 10.5937/bezbednost2402005P.

[20] S. J. Nightingale and K. A. Wade, "Identifying and minimising the impact of fake visual media: Current and future directions," *Memory, Mind & Media*, vol. 1, p. e15, Oct. 2022, doi: 10.1017/mem.2022.8.

[21] A. Mishra, A. Bharwaj, A. K. Yadav, K. Batra, and N. Mishra, "Deepfakes - Generating Synthetic Images, and Detecting Artificially Generated Fake Visuals Using Deep Learning," in *2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, IEEE, Jan. 2024, pp. 587–592. doi: 10.1109/Confluence60223.2024.10463337.

[22] S. Ahmed, "Navigating the maze: Deepfakes, cognitive ability, and social media news skepticism," *New Media Soc*, vol. 25, no. 5, pp. 1108–1129, May 2023, doi: 10.1177/14614448211019198.

[23] J. Roozenbeek and S. van der Linden, "Fake news game confers psychological resistance against online misinformation," *Palgrave Commun*, vol. 5, no. 1, p. 65, Jun. 2019, doi: 10.1057/s41599-019-0279-9.

[24] A. Sanchez-Acedo, A. Carbonell-Alcocer, M. Gertrudix, and J.-L. Rubio-Tamayo, "The challenges of media and information literacy in the artificial intelligence ecology: deepfakes and misinformation," *Communication & Society*, pp. 223–239, Oct. 2024, doi: 10.15581/003.37.4.223-239.

[25] S. Shawky and E. Mokadem, “The Effect of Media Literacy on Misinformation and Deep Fake Video Detection,” *Arab Media & Society*, no. 35, 2023.
[26] A. Birrer and N. Just, “What we know and don’t know about deepfakes: An investigation into the state of the research and regulatory landscape,” *New Media Soc*, May 2024, doi: 10.1177/14614448241253138.
[27] R. Wan *et al.*, “Community-driven AI: Empowering people through responsible data-driven decision-making,” in *Computer Supported Cooperative Work and Social Computing*, New York, NY, USA: ACM, Oct. 2023, pp. 532–536. doi: 10.1145/3584931.3611282.
[28] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, “Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement,” *BMJ*, vol. 339, no. jul21 1, pp. b2535–b2535, Jul. 2009, doi: 10.1136/bmj.b2535.
[29] J. Thomas and A. Harden, “Methods for the thematic synthesis of qualitative research in systematic reviews,” *BMC Med Res Methodol*, vol. 8, no. 1, p. 45, Dec. 2008, doi: 10.1186/1471-2288-8-45.