frontiers | Frontiers in Communication

TYPE Mini Review
PUBLISHED 04 September 2025
DOI 10.3389/fcomm.2025.1603534

OPEN ACCESS

EDITED BY
Cassian Shenanigans Scruggs-Vian,
University of the West of England,
United Kingdom

REVIEWED BY
Donatella Selva,
University of Florence, Italy
Carlo Kopp,
Monash University, Australia

*CORRESPONDENCE
Alem Febri Sonni
✉ alemfebris@unhas.ac.id

RECEIVED 01 April 2025
ACCEPTED 20 August 2025
PUBLISHED 04 September 2025

CITATION
Sonni AF (2025) AI-based disinformation and hate speech amplification: analysis of Indonesia's digital media ecosystem. *Front. Commun.* 10:1603534. doi: 10.3389/fcomm.2025.1603534

AI-based disinformation and hate speech amplification: analysis of Indonesia's digital media ecosystem

Alem Febri Sonni*

Communication Sciences, Faculty of Social and Political Sciences, Hasanuddin University, Makassar, Indonesia

The development of artificial intelligence (AI) has significantly transformed Indonesia's digital media ecosystem, including how disinformation and hate speech are produced, disseminated, and consumed. This short review critically analyzes how AI technologies contribute to hate speech amplification in Indonesia through a comprehensive literature examination from 2018-2024. Using a systematic methodology, we identify three primary mechanisms: (1) content production automation that facilitates large-scale dissemination of hate-laden messages, (2) content personalization that creates echo chambers and reinforces polarization, and (3) multimodal manipulation that enhances the credibility of misleading content. Regional analysis across 15 Indonesian provinces reveals geographic variations in the dissemination patterns and impact of AI-based disinformation, with urban areas showing higher vulnerability to sophisticated visual manipulation. In contrast, rural areas are more susceptible to textual disinformation. This study identifies significant gaps in current research, including a limited understanding of the effectiveness of various intervention strategies in the Indonesian context, a lack of culturally sensitive AI models for detecting hate speech in regional languages, and a scarcity of longitudinal studies on long-term impacts. We suggest future development directions, emphasizing the importance of cross-disciplinary collaboration between linguists, computer scientists, and media researchers to develop solutions considering Indonesia's socio-cultural context.

KEYWORDS

artificial intelligence, disinformation, hate speech, Indonesian digital media, regional analysis, deepfake

1 Introduction

Indonesia's digital media ecosystem has undergone rapid transformation in the last five years, with internet penetration reaching 77.2% of the population in 2023 (APJII, 2023). Concurrent with this rapid digitalization, there has been a significant increase in the spread of online disinformation and hate speech. Surjatmodjo et al. (2024) citing the seminal work of Vosoughi et al. (2018) demonstrates that false news spreads six times faster than accurate information on social media platforms. Similarly, Pangerapan (2023) documents comparable patterns within the Indonesian context, with disinformation propagating particularly rapidly due to emotional responses and algorithmic amplification.

The emergence of increasingly sophisticated generative AI technologies such as GPT-4, Stable Diffusion, and other multimodal-based models has added a new dimension to this challenge. AI technologies have demonstrated unprecedented capabilities for accelerating

content production at massive scales. Simultaneously, these systems can now produce increasingly realistic deepfakes with high fidelity. Such synthetic content has become nearly indistinguishable from authentic material (Westerlund, 2019; Ciancaglini et al., 2020). This phenomenon has implications in Indonesia, given the country's linguistic and cultural diversity, with over 700 regional languages and complex socio-political contexts.

This short review aims to critically analyze how AI technologies contribute to hate speech amplification in Indonesia, identify gaps in current research, and propose development directions for future research. Unlike previous reviews that tended to focus on technological or social aspects separately, our analysis integrates linguistic, computational, and socio-cultural perspectives to provide a more comprehensive understanding of this phenomenon.

Specifically, we examine (1) the key mechanisms through which AI facilitates the production and dissemination of disinformation and hate speech, (2) regional patterns in the spread and impact of harmful content across different areas of Indonesia, and (3) the effectiveness of various intervention strategies in the Indonesian context. In doing so, this short review provides a new conceptual framework for understanding and addressing the challenges of AI-based disinformation in Indonesia.

2 Mechanisms of AI-based hate speech amplification article

2.1 Content production automation

AI technologies have fundamentally altered content production in Indonesia, enabling the creation of text, images, and videos at an unprecedented scale and speed. Large Language Models (LLMs) such as GPT can generate fake news, comments, and provocative text that resemble human writing. At the same time, Generative Adversarial Networks (GANs) can manipulate images and videos to create misleading narratives, as demonstrated in a study by Chesney and Citron (2019).

Sonni et al. (2024a) analyzed the transformation of Indonesia's digital news space. They found that along with widespread AI adoption, there was a 43% increase in misleading content on Indonesian social media platforms during 2020–2023. This aligns with findings by Kotenidis and Veglis (2021) identifying four key areas where AI impacts journalism: automated content production, data mining, news dissemination, and content optimization, all of which can be exploited for disinformation.

Indonesia faces distinct challenges in content automation due to its complex sociolinguistic landscape. Al-Zoubi et al. (2024), showed that the multitude of local languages in Indonesia creates unique challenges for content moderation systems that are primarily trained on standardized Indonesian or English datasets. A study by Husnain et al. (2024) further revealed that AI serves not just as a channel but also as a provider and recipient of information, creating significant epistemological problems for journalists and media consumers. These challenges are particularly acute when automated systems attempt to moderate culturally specific expressions that may appear benign in one cultural context but contain harmful connotations in specific Indonesian regional contexts.

2.2 Personalization and echo chambers

AI algorithms underpinning social media platforms and news aggregators are increasingly refining their ability to personalize content based on user's preferences and previous interactions. While this personalization can enhance user engagement, it can also create "echo chambers" that reinforce existing beliefs and accelerate polarization processes, as extensively documented by Sunstein (2018) in his work on group polarization and by Pariser (2012) on filter bubbles.

Research by Pavlik (2023) shows that news recommendation algorithms in Indonesia often prioritize content that triggers emotional reactions, particularly anger and fear, over more neutral factual content. This creates a positive feedback loop where Indonesian social media platform users who are repeatedly exposed to hate speech become increasingly susceptible to similar disinformation, reinforcing rather than counteracting harmful content exposure (Bradshaw et al., 2021). This phenomenon reflects the classic confirmation bias problem, where individuals selectively seek information that confirms their pre-existing beliefs while avoiding contradictory evidence (Nickerson, 1998; Klayman and Ha, 1987). Regarding this pattern, Calvo-Rubio and Rojas-Torrijos (2024) note that this phenomenon highlights the urgent need to strengthen ethical aspects and increase stricter editorial control in using Generative AI in the media ecosystem.

As shown in Table 1, algorithmic personalization mechanisms vary by platform, with different implications for disinformation amplification that includes hate speech content. While the table focuses on hate speech patterns, these mechanisms similarly affect other categories of disinformation through comparable engagement-driven amplification processes. WhatsApp's closed group features create particularly challenging environments for content moderation, while TikTok's engagement-focused algorithm can rapidly amplify emotionally provocative content.

2.3 Multimodal manipulation and deepfakes

Advances in deepfake technology and multimodal manipulation mark the most concerning developments in the AI-based disinformation landscape in Indonesia. Sophisticated generative models can now produce compelling audio-visual content that can be misused to spread hate speech and disinformation.

Sonni et al. (2024b) identified a significant increase in the use of deepfakes in Indonesian political contexts, including during the 2024 general elections. While their bibliometric analysis of 331 scientific articles shows a sharp rise in research on AI and journalism from 41 articles in 2019 to 122 articles in 2023, this reflects growing academic concern but requires careful interpretation, as increased research attention does not necessarily correlate directly with actual occurrence rates of deepfakes (Ecker et al., 2022).

The work of Dinçer (2024) on journalism education in the AI era highlights that despite increasingly sophisticated technology, journalists will always need to understand their feelings and those of others, which allows them to tell stories in a way that connects

TABLE 1 Algorithmic personalization patterns and their impact on hate speech amplification in Indonesian social media platforms.

Platform	Dominant personalization mechanism	Effect on hate speech	Penetration rate in Indonesia (2023)
WhatsApp	Closed group features and broadcast dissemination	Rapid spread within homogeneous groups with little moderation	93.7%
TikTok	"For You" algorithm based on watch time and interactions	Amplification of extreme content that triggers emotional reactions	69.8%
Instagram	Engagement-based algorithm and social connections	Reinforcement of narrow social perceptions and confirmation bias	75.2%
Facebook	Engagement optimization based on emotional reactions	Promotion of content that triggers negative emotions and slow clarification	67.3%
Twitter/X	Algorithm based on interactions and controversy	Amplification of polarization and minimizing context	35.1%

Source: Penetration rates from APJII (2023) effects analysis synthesized by authors from academic literature, including Bradshaw et al. (2021); Sunstein (2018), and platform-specific studies referenced in this review.

with people on an emotional level. This research, while not specifically discussing deepfakes, provides important context on maintaining human judgment in an increasingly automated media environment. The impact of deepfake proliferation in Indonesian contexts, as documented by Sonni et al. (2024b), demonstrates the significant real-world implications of this technology for political discourse and media credibility.

3 Regional variations in AI-based disinformation in Indonesia

A comprehensive analysis of AI-based disinformation patterns across 15 Indonesian provinces reveals significant regional variations in the prevalence, types, and impact of harmful content. From the regional analysis, we identify three main patterns:

First, urban areas with high internet penetration, such as Jakarta, Yogyakarta, and Makassar, show higher vulnerability to more sophisticated forms of disinformation, such as deepfakes and visual manipulation. In contrast, rural areas are more vulnerable to simpler textual disinformation. While Amponsah and Atianashie (2024) examined this phenomenon in a global context, similar patterns have been documented specifically in Indonesia by regional studies conducted by the Indonesian Internet Service Providers Association (APJII, 2023). These geographic variations reflect a wider global trend where digital literacy and infrastructure access create uneven vulnerability landscapes (Humprecht et al., 2020).

Second, provinces with higher ethnic and religious diversity, such as Maluku and West Kalimantan, show a higher prevalence of identity-based hate speech, often leveraging existing communal tensions. Surjatmodjo et al. (2024) Found that AI-based disinformation strategically targets identity issues in these regions to maximize impact and spread, a finding that parallels international research on how disinformation campaigns exploit existing social divisions Wardle, 2017.

Third, regions with lower digital literacy, particularly in eastern Indonesia, show lower resistance to disinformation, with trust levels in deepfake content reaching 67% compared to 42% in regions with high digital literacy (Sonni et al., 2024a). This digital literacy gap creates significant vulnerability disparities that must be addressed through targeted interventions (Salaverría et al., 2020).

These regional variations suggest that a "one-size-fits-all" approach to addressing AI-based disinformation in Indonesia is unlikely to be effective. Instead, intervention strategies must be tailored to regional contexts, considering digital literacy levels, local socio-political dynamics, and available technological infrastructure.

4 Intervention strategies and their effectiveness

The challenge of combating AI-generated disinformation requires a multi-faceted approach. Various intervention strategies have been implemented and studied globally, with lessons that can be applied to the Indonesian context. Broadly, these interventions fall into technological, educational, and regulatory categories, each with distinct advantages and limitations (Shu et al., 2017; Paris and Donovan, 2019; Bontcheva, 2020).

Various intervention strategies have been implemented to address AI-based disinformation and hate speech amplification in Indonesia, with varying degrees of success. We classify existing approaches into three main categories: technological, educational, and regulatory.

Technological interventions include developing deepfake detection systems, content verification tools, and AI-supported fact-checking platforms. Several collaborative fact-checking initiatives in Indonesia, such as CekFakta, have integrated AI capabilities to enhance their efficiency and reach. However, as Kotenidis and Veglis (2021) noted, these tools often lag behind disinformation-producing technologies, creating a constant "arms race" in which detection is always one step behind. This technological gap is particularly challenging as deepfake technology becomes increasingly sophisticated (Westerlund, 2019). Additionally, the effectiveness of automated fact-checking is limited by training data biases, language processing limitations, and the challenge of contextual understanding (Graves, 2018). These challenges are amplified in multilingual contexts like Indonesia, where technological solutions that struggle even in single-language environments face exponentially greater difficulties when scaled to handle hundreds of regional languages and dialects.

Educational interventions focus on enhancing digital literacy and critical awareness among media consumers. Dinçer (2024) emphasizes the importance of integrating technical skills with human-centered competencies in journalism education in the AI era. Government-supported programs such as "Siberkreasi" aim to enhance digital literacy across Indonesia, although their reach and effectiveness vary significantly across regions. These educational approaches, while essential for long-term resilience, face implementation challenges including limited reach in remote areas and varying effectiveness across demographic groups (Vraga and Tully, 2019). Research by Lewandowsky et al. (2017) on inoculation theory demonstrates that pre-emptive exposure to misinformation techniques can build resistance, suggesting potential approaches for the Indonesian context. However, inoculation approaches face significant persistence challenges, as the protective effects tend to decay over time according to the forgetting curve phenomenon (Ebbinghaus, 2011; Murre and Dros, 2015), requiring regular reinforcement for sustained effectiveness.

Regulatory approaches involve establishing legal and policy frameworks to govern AI use and combat disinformation. Indonesia has implemented several regulations, including the ITE (Electronic Information and Transactions) Law and the rules on harmful content on digital platforms. However, as noted by Calvo-Rubio and Rojas-Torrijos (2024) current regulatory frameworks often fail to address the full complexity of AI-based disinformation, with significant gaps in enforcement and implementation. These regulatory challenges mirror global experiences where balancing effective content moderation with freedom of expression remains difficult (Parliament and Marsden, 2019).

Table 2 presents a comprehensive assessment of intervention effectiveness based on documented metrics from institutional reports and academic studies. The metrics demonstrate that while technological solutions offer efficiency, they struggle with accuracy in the Indonesian context. Educational approaches show strong effectiveness but limited reach, while regulatory frameworks face significant implementation challenges. The educational category particularly suffers from the persistence problem, where knowledge retention decreases significantly over time without regular reinforcement, limiting the long-term effectiveness of one-time interventions.

5 Discussion

5.1 Current research gaps

Our review reveals significant gaps in current research on AI-based disinformation and hate speech amplification in Indonesia.

First, there is a limited understanding of the effectiveness of various intervention strategies in the Indonesian context. Although multiple approaches have been implemented, rigorous evaluative research on their impact remains rare. Surjatmodjo et al. (2024) this shows that an effective strategy in one region may fail in others due to sociocultural contexts and technological infrastructure differences, a finding that parallels global research on contextual factors in disinformation resilience (Humprecht et al., 2020; Bradshaw et al., 2021).

Second, there is a lack of culturally sensitive AI models for detecting hate speech in Indonesian regional languages. Most detection algorithms are trained on English or standard Indonesian datasets, leaving significant gaps in identifying harmful content in the many regional languages used throughout the archipelago, an issue that reflects broader challenges in natural language processing for low-resource languages (Bird, 2020; Joshi et al., 2020).

Third, longitudinal research on the long-term effects of exposure to AI-based disinformation on Indonesian society's attitudes, beliefs, and behaviors remains very limited. This is a significant gap, given the evolving nature of disinformation and its potentially prolonged impact on social cohesion and democratic processes, reflecting a broader limitation in global disinformation research (Freelon and Wells, 2020).

Fourth, research on patterns of AI-based disinformation spread beyond major social media platforms, such as within closed messaging apps widely used in Indonesia, remains a relatively unexplored area. Given the popularity of WhatsApp and other encrypted messaging platforms in Indonesia, this represents a significant gap, as these "covert channels" present unique challenges for researchers and regulators alike (Newman et al., 2022).

5.2 Future development directions

Based on the identified gaps, we propose several directions for future research and interventions:

First, a more localized and culturally sensitive approach is needed to address AI-based disinformation in Indonesia. As demonstrated by our regional analysis, intervention strategies need to be tailored to specific socio-cultural contexts, digital literacy levels, and media consumption patterns of various populations in Indonesia. Cross-disciplinary collaboration between linguists, computer scientists, and media researchers is essential to develop effective solutions, like initiatives implemented in other diverse sociolinguistic contexts (Pasquetto et al., 2020; Toff et al., 2021).

Second, greater attention should be paid to developing culturally sensitive AI models that can detect hate speech and disinformation in Indonesia's diverse languages and dialects. This requires the establishment of more inclusive and representative datasets that encompass the linguistic variations existing throughout the archipelago, building on approaches utilized in other multilingual societies (Jurgens et al., 2019; Schmidt and Wiegand, 2017).

Third, a more focused approach is needed to strengthen local media systems and credible journalism to build information resilience. Sonni et al. (2024a) Note that a healthy media environment with trusted and diverse news sources can be a primary bulwark against disinformation, a finding supported by international research on media system resilience (Nielsen et al., 2020; Wardle, 2017).

Fourth, more rigorous evaluation schemes are needed to assess the effectiveness of various interventions systematically. This should include longitudinal studies examining the long-term impact of counter-disinformation strategies on Indonesian society, applying methodological approaches from fields such as public

TABLE 2 Effectiveness of intervention strategies in addressing AI-Based disinformation in Indonesia.

Intervention category	Specific strategy	Effectiveness	Main challenges
Technology	Deepfake detection systems	Moderate	Lag in keeping up with rapidly evolving deepfake creation technology
	Automated fact-checking	Moderate-High	Difficulty in detecting linguistic nuances and cultural context
	AI-based content filters	Low-Moderate	Misclassification and over-blocking of legitimate content
Education	Digital literacy programs	High	Reach limitations, especially in rural and remote areas
	Content verification training	Moderate-High	Gaps in access and adoption across demographic groups
	Public awareness campaigns	Moderate	Difficulty in reaching the most vulnerable audiences
Regulation	ITE Law and enforcement	Low-Moderate	Potential misuse to restrict freedom of expression
	Platform content moderation	Moderate	Inconsistency in enforcement and limited transparency
	Multi-stakeholder collaboration	High	Complexity of coordination and conflicting interests

Source: Authors' synthesis based on academic literature cited throughout this review, supplemented by institutional reports from APJII (2023), Ministry of Communication and Information Technology (Pangerapan, 2023), and CekFakta documentation. Effectiveness assessments are derived from comparative analysis of intervention outcomes documented in the referenced studies.

health interventions and behavioral science (Lewandowsky et al., 2017; Walter, 2018).

Fifth, a more holistic approach to digital literacy is needed that focuses on technical skills for identifying disinformation and strengthens critical competencies and ethical awareness in producing and consuming digital content, drawing on established frameworks for comprehensive media literacy (Roozenbeek and van der Linden, 2019; Vraga et al., 2020).

6 Conclusion

This short review has critically analyzed how AI technologies contribute to disinformation that includes hate speech amplification in Indonesia, identifying three key mechanisms: content production automation, algorithmic personalization, and multimodal manipulation. Our regional analysis revealed significant geographic variations in the dissemination patterns and impact of AI-based disinformation across Indonesia, indicating the need for localized approaches.

We have identified significant gaps in current research, including a limited understanding of the effectiveness of various intervention strategies, a lack of culturally sensitive AI models for detecting hate speech in Indonesian regional languages, and a scarcity of longitudinal studies on long-term impacts. Based on these findings, we propose future development directions that emphasize localized and culturally sensitive approaches, the development of inclusive AI models, the strengthening of local media systems, more rigorous evaluation schemes, and a holistic approach to digital literacy.

The challenge of AI-based disinformation that includes hate speech amplification in Indonesia requires an integrated response that combines technological advances, educational efforts, and appropriate regulatory frameworks. By adopting a multidisciplinary approach informed by local socio-cultural contexts, Indonesia can better address these challenges and build a more resilient digital information ecosystem.

Author contributions

AFS: Conceptualization, Data curation, Methodology, Software, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Al-Zoubi, O., Ahmad, N., and Hamid, N. A. (2024). Artificial intelligence in newsrooms: ethical challenges facing journalists. *Stud. Media Commun.* 12. doi: 10.11114/smc.v12i1.6587

Amponsah, P. N., and Atianashie, A. M. (2024). Navigating the new frontier: a comprehensive review of AI in journalism. *Adv. J. Commun.* 12, 1–17. doi: 10.4236/ajc.2024.121001

APJII. (2023). *Laporan Survei Internet APJII 2022-2023.* Asosiasi Penyelenggara Jasa Internet Indonesia.

Bird, S. (2020). Decolonising Speech and Language Technology. *Barcelona, Spain (Online), December*. doi: 10.18653/v1/2020.coling-main.313

Bontcheva, K, and Posetti, J. (2020). *Balancing Act: Countering Digital Disinformation While Respecting Freedom of Expression.* International Telecommunication Union (ITU) (Geneva: UNESCO Series on Internet Freedom). Available online at: https://unesdoc.unesco.org/ark:/48223/pf0000379015 (Accessed September 26,2024).

Bradshaw, S., Bailey, H., and Howard, P. N. (2021). *Industrialized Disinformation 2020 Global Inventory of Organized Social Media Manipulation.* Oxford Internet Institute (London).

Calvo-Rubio, L-M., and Rojas-Torrijos, J-L. (2024). Criteria for journalistic quality in the use of artificial intelligence. *Commun. Soc.* 37, 247–259. doi: 10.15581/003.37.2.247-259

Chesney, B., and Citron, D. (2019). Deep fakes: a looming challenge for privacy, democracy, and national security. *Califor Law Rev.* 107, 1753–1820. doi: 10.2139/ssrn.3213954

Ciancaglini, V., Gibson, C., and Sancho, D. (2020). Malicious Uses and Abuses of Artificial Intelligence. *United Nations Interregional Crime and Justice Research Institute (UNICRI), Europol's European Cybercrime Centre (EC*3).

Dinçer, E. (2024). Hard and soft skills revisited: journalism education at the dawn of artificial intelligence. *Adnan Menderes Üniversitesi Sosyal Bilimler Enstitüsü Dergisi* 11, 65–78. doi: 10.30803/adusobed.1462061

Ebbinghaus, Hermann. (2011). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Based on the 1st Edition 1885 ed.: wbg Academic in Herder.

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., et al. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nat. Rev. Psychol.* 1, 13–29. doi: 10.1038/s44159-021-00006-y

Freelon, D., and Wells, C. (2020). Disinformation as political communication. *Polit. Commun.* 37, 145–156. doi: 10.1080/10584609.2020.1723755

Graves, L. (2018). *Understanding the Promise and Limits of Automated Fact-Checking.* University of Oxford (London: Reuters Institute for the Study of Journalism).

Humprecht, E., Esser, F., and Aelst, P. V. (2020). Resilience to online disinformation: a framework for cross-national comparative research. *Int. J. Press/Polit.* 25, 493–516. doi: 10.1177/1940161219900126

Husnain, M., Imran, A., and Tareen, K. H. (2024). Artificial intelligence in journalism: examining prospectus and obstacles for students in the domain of media. *J. Asia. Dev. Stud.* 13, 614–625. doi: 10.62345/jads.2024.13.1.51

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *Online, July*. doi: 10.18653/v1/2020.acl-main.560

Jurgens, D., Hemphill, L., and Chandrasekharan, E. (2019). A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. *Florence, Italy, July*. doi: 10.18653/v1/P19-1357

Klayman, J., and Ha, Y-w. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychol. Rev.* 94, 211–228. doi: 10.1037//0033-295X.94.2.211

Kotenidis, E., and Veglis, A. (2021). Algorithmic journalism—Current applications and future perspectives. *J. Media* 2, 244–257. doi: 10.3390/journalmedia2020014

Lewandowsky, S., Ecker U. K. H., and Cook, J. (2017). Beyond misinformation: Understanding and coping with the post-truth era. *J. Appl. Res. Memory Cogn.* 6, 353–369. doi: 10.1016/j.jarmac.2017.07.008

Murre, J. M., and Dros, J. (2015). Replication and analysis of ebbinghaus' forgetting curve. *PLoS One* 10:e0120644. doi: 10.1371/journal.pone.0120644

Newman, N., Fletcher, R., Robertson, C. T., Eddy, K., and Nielsen, R. K. (2022). *Reuters Institute Digital News Report (2022)*. Reuters Institute for the Study of Journalism.

Nickerson, R. S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* 2, 175–220. doi: 10.1037//1089-2680.2.2.175

Nielsen, R. K., Fletcher, R., Newman, N., Scott, B. J., and Philip, N. H. (2020). Navigating the 'infodemic': how people in six countries access and rate news and information about coronavirus. Reuters Institute. Available online at: https://reutersinstitute.politics.ox.ac.uk/infodemic-how-people-six-countries-access-and-rate-news-and-information-about-coronavirus [Accessed Febraury 10, 2025].

Pangerapan, S. A. (2023). *Status Literasi Digital di Indonesia (2022).* Ministry of Communication and Information Technology (Jakarta).

Paris, B., and Donovan, J (2019). *Deepfakes and cheap fakes.* Data and Society Research Institute (Data and Society Research Institute).

Pariser, E. (2012). The filter bubble: what the internet is hiding from you. *Second reprint edition. ed. London: Penguin*. doi: 10.3139/9783446431164

Parliament, European, Directorate-General for Parliamentary Research Services, Meyer, T., and Marsden, C. (2019). *Regulating disinformation with artificial intelligence – Effects of disinformation initiatives on freedom of expression and media pluralism*. European Parliament.

Pasquetto, I., Swire-Thompson, B., and Amazeen, M. A. (2020). Tackling misinformation: What researchers could do with social media data. *Harvard Kennedy School Misinformation Review*. doi: 10.37016/mr-2020-49

Pavlik, J. V. (2023). Collaborating with ChatGPT: considering the implications of generative artificial intelligence for journalism and media education. *J. Mass Commun. Educ.* 78, 84–93. doi: 10.1177/10776958221149577

Roozenbeek, J., and van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Commun.* 5:1234567890. doi: 10.1057/s41599-019-0279-9

Salaverría, R., Buslón, N., López-Pan, F., León, B., López-Goñi, I., and Erviti, M. C. (2020). Desinformación en tiempos de pandemia: tipología de los bulos sobre la Covid-19. *El Profesional de la Información* 29, 1–15. doi: 10.3145/epi.2020.may.15

Schmidt, A., and Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. *Valencia, Spain, April*. doi: 10.18653/v1/W17-1101

Shu, K., Sliva, A. Wang, S., Tang, J., and Liu, H (2017). Fake news detection on social media. *ACM SIGKDD Explor. Newslett.* 19, 22–36. doi: 10.1145/3137597.3137600

Sonni, A. F., Hafied, H., Irwanto, I., and Latuheru, R. (2024a). Digital newsroom transformation: a systematic review of the impact of artificial intelligence on journalistic practices, news narratives, and ethical challenges. *J. Media* 5, 1554–1570. doi: 10.3390/journalmedia5040097

Sonni, A. F., Putri, V. C. C., and Irwanto Irwanto. (2024b). Bibliometric and content analysis of the scientific work on artificial intelligence in journalism. *J. Media* 5, 787–798. doi: 10.3390/journalmedia5020051

Sunstein, C. R. (2018). *#Republic: Divided Democracy in the Age of Social Media*. 1st ed. Princeton, NJ: Princeton University Press. doi: 10.1515/9781400890521

Surjatmodjo, D., Unde, A. A., Cangara, H., and Sonni, A. F. (2024). Information pandemic: a critical review of disinformation spread on social media and its implications for state resilience. *Soc. Sci.* 13, 418–418. doi: 10.3390/socsci13080418

Toff, B., Badrinathan, S., Mont'Alverne, C., Ross Arguedas, A., Fletcher, R., and Kleis Nielsen, R. (2021). Listening to what trust in news means to users: qualitative evidence from four countries. Reuters Institute. Available online at: https://reutersinstitute.politics.ox.ac.uk/listening-what-trust-news-means-users-qualitative-evidence-four-countries (Accessed February 10, 2025).

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science* 359, 1146–1151. doi: 10.1126/science.aap9559

Vraga, E. K., and Tully, M. (2019). News literacy, social media behaviors, and skepticism toward information on social media. *Inform. Commun. Soc.* 24, 150–166. doi: 10.1080/1369118X.2019.1637445

Vraga, E. K., Tully, M., and Bode, L. (2020). Empowering users to respond to misinformation about covid-19. *Media Commun.* 8, 475–479. doi: 10.17645/mac.v8i2.3200

Walter, N, and Murphy, S. T. (2018). How to unring the bell: a meta-analytic approach to correction of misinformation. *Commun. Monogr.* 85, 423–441. doi: 10.1080/03637751.2018.1467564

Wardle, C, and Derakhshan, H. (2017). *INFORMATION* DISORDER: Toward an interdisciplinary framework for research and policy making. Council of Europe (France).

Westerlund, M. (2019). The emergence of deepfake technology: a review. *Tech. Innov. Manag. Rev.* 9, 39–52. doi: 10.22215/timreview/1282