Review EL'BAHITH - ENS - Bouzareah - Algiers					
ISSN : 9577-1112	EISSN : 2602-5388	Volume: (18)	Number : (02)	year: 2026	Pages: 278-293

CREDIBILITY OF SOCIAL MEDIA IN THE AGE OF ARTIFICIAL INTELLIGENCE: A CRITICAL AND COMPARATIVE APPROACH TO DIGITAL TRUST

مصداقية وسائل التواصل الاجتماعي في عصر الذكاء الاصطناعي: مقاربة نقدية ومقارنة للثقة الرقمية

Nasreddine HADID
University of Algiers 3
nasreddinehadid@gmail.com
Hakima JABALLAH
University of Algiers 3
hchenit@yahoo.fr

Date of receipt: 25/04/2026 *Date of acceptance: 04/06/ 2026* *Date of publication: 10/06/2026*

Corresponding Author :Nasreddine Hadid, nasreddinehadid@gmail.com

ABSTRACT

This article examines the credibility crisis of social media in the age of artificial intelligence through a critical and comparative approach to digital trust. It argues that credibility has shifted from an institutional attribute to a contested outcome shaped by users, platforms, algorithms, and AI systems. Drawing on recent academic studies, the article analyzes how algorithmic amplification, generative AI, and deepfakes intensify misinformation, while also exploring the verification potential of AI-enabled tools. The discussion compares electoral, health, disaster, and gendered online contexts to show that digital trust is shaped by local vulnerabilities, platform design, and institutional accountability. The article concludes that sustainable digital trust requires a hybrid model combining AI-assisted detection, human editorial judgment, transparent governance, and context-sensitive media literacy.

Keywords:

Artificial intelligence; Deepfakes; Digital trust; Misinformation; Social media credibility.

Review EL'BAHITH - ENS - Bouzareah - Algiers					
ISSN : 9577-1112	EISSN : 2602-5388	Volume: (18)	Number : (02)	year: 2026	Pages: 278-293

ملخص

تتناول هذه الدراسة أزمة مصداقية وسائل التواصل الاجتماعي في عصر الذكاء الاصطناعي من خلال مقاربة نقدية ومقارنة لمفهوم الثقة الرقمية. وتنطلق من أن المصداقية لم تعد خاصية مؤسساتية ثابتة، بل أصبحت نتيجة تفاوضية تتشكل بفعل تفاعل المستخدمين والمنصات والخوارزميات وأنظمة الذكاء الاصطناعي. وبالاستناد إلى دراسات أكاديمية حديثة، تحلل الدراسة أثر التضخيم الخوارزمي والذكاء الاصطناعي التوليدي وتقنيات التزييف العميق في تعميق المعلومات المضللة، مع إبراز الإمكانات التي تتيحها أدوات التحقق المدعومة بالذكاء الاصطناعي.

وتخلص الدراسة إلى أن بناء ثقة رقمية مستدامة يقتضي نموذجا هجينا يجمع بين الكشف الآلي والمراجعة البشرية والحوكمة الشفافة والتربية الإعلامية الحساسة للسياق.

الكلمات المفتاحية:

التزييف العميق؛ الثقة الرقمية؛ الذكاء الاصطناعي؛ المعلومات المضللة؛ مصداقية وسائل التواصل الاجتماعي.

1.Introduction

The question of credibility of social media platforms has become one of the most urgent issues in the contemporary digital environment. The expansion of networked platforms has intensified the circulation of misleading claims, fabricated stories, and emotionally manipulative visuals, often at a speed that surpasses the capacity of journalism, educational institutions, and regulatory actors to verify and contextualize them. Tandoc et al. (2018) show that the term fake news is used inconsistently across scholarship and often covers several forms of deceptive or misleading content. In the platform era, this conceptual instability is compounded by economic and algorithmic conditions that reward visibility and rapid circulation.

Credibility has therefore shifted from being a stable institutional attribute to becoming a contested outcome of interactions among users, platforms, algorithms, and increasingly artificial intelligence systems. In many societies, including North Africa and the broader Maghreb, social media has become a primary source of information, interpretation, and public discussion, sometimes surpassing traditional outlets in immediacy and everyday relevance. The result is a widening crisis of digital trust, understood here as the degree to which audiences believe that the information, interfaces, and recommendation systems they encounter online are dependable, accountable, and socially legitimate.

Artificial intelligence occupies a paradoxical position within this crisis. On the one hand, recent advances in generative AI have lowered the cost of producing plausible text, audio, images, and video, thereby expanding the scale of synthetic misinformation and deepfakes (World Economic Forum, 2024). On the other hand, AI also powers advanced detection systems, multilingual monitoring tools, and forensic methods that can assist verification. The central question is thus not whether AI is inherently trustworthy or untrustworthy, but under what institutional, legal, professional, and cultural conditions AI can contribute to restoring digital trust rather than destabilizing it further.

This article revisits credibility in the age of AI through a critical and comparative lens. It combines conceptual discussion with recent empirical findings, integrates lived examples from electoral, health, disaster, and gendered online environments, and argues for a hybrid model in which algorithmic assistance remains subordinate to human accountability, transparent governance, and context-sensitive media literacy.

2.Study Problem

The main research question guiding this article is: To what extent can artificial intelligence strengthen the credibility of social media content, and under what conditions does it instead deepen the ongoing crisis of digital trust?

This primary question branches into several sub-questions: How should credibility be theorized in algorithmically curated environments? How do deepfakes and AI-generated texts alter public perceptions of truth and authenticity? What do recent studies conducted in different social settings reveal about credibility judgments? And what regulatory, educational, and professional arrangements are most capable of rebuilding trust without undermining freedom of expression?

3.Theoretical and Conceptual Framework of Digital Credibility and Trust

3.1. Credibility as a Cognitive, Social, and Algorithmic Judgment

In classic communication research, credibility referred largely to source expertise and trustworthiness. In digital environments, however, users encounter overwhelming quantities of content produced by institutions, influencers, ordinary users, bots, and generative systems. Metzger and Flanagin (2013) argue that under such conditions people often rely on cognitive heuristics rather than systematic verification. Interface cues, emotional tone, repetition, popularity metrics, and perceived familiarity can all substitute for genuine evidence.

Yet credibility is never only cognitive. It is also socially constructed. Charaudeau (2011) treats media credibility as bound up with discourse, public expectations, and negotiated legitimacy rather than as a purely technical property of information. What counts as believable in one community may not be persuasive in another, because credibility depends on shared norms, trust histories, symbolic authority, and social experience. This is especially important in societies marked by institutional distrust, political polarization, or uneven access to formal news infrastructures.

A third dimension is algorithmic credibility. Visibility itself is often mistaken for validity. When platforms repeatedly surface certain posts, users may infer that these posts have already passed some threshold of relevance or reliability, even though ranking systems usually optimize for engagement rather than truthfulness. A recent systematic review shows that algorithmic priorities increasingly shape newsroom routines as well, pulling editorial decisions toward content likely to circulate and be shared rather than content most conducive to public understanding (Hastuti et al., 2025).

3.2. Digital Trust as a Systemic Condition

Digital trust is broader than the credibility of any single post. It includes confidence in the platform, in the fairness of recommendation systems, in the integrity of verification mechanisms, and in the accountability of institutions that mediate online communication. Tsfati and Cappella (2003) remind us that trust always includes vulnerability: users depend on information they cannot fully verify on their own. When opaque infrastructures repeatedly expose them to manipulation, this vulnerability becomes politically and psychologically costly.

From the perspective of network society theory, the trust crisis is structural rather than accidental. Users inhabit dense informational networks in which traditional markers of authority are weakened while algorithmic gatekeepers remain largely invisible. The result is not simply the presence of more falsehoods, but the erosion of stable epistemic reference points. In such an environment, even true information can lose persuasive power because the very conditions under which truth is evaluated have been destabilized.

4. Social Media, Misinformation, and the Deepfake Turn

Before examining algorithmic amplification and generative AI separately, it is necessary to situate misinformation within a broader transformation of platform credibility. Social media no longer functions merely as a channel for the circulation of false claims; it has become an environment in which visibility, emotion, automation, and synthetic media interact to redefine what users perceive as credible. The deepfake turn is therefore not only a technical development in image, audio, and video manipulation, but also a communicative shift that weakens traditional cues of authenticity and intensifies uncertainty around public evidence.

4.1. Algorithmic Amplification and the Attention Economy

The economics of digital platforms reward velocity, novelty, outrage, and affective intensity. Vosoughi et al. (2018) famously demonstrated that false news

diffuses farther, faster, and more broadly than true news on Twitter, not because bots alone drive the process, but because humans are especially responsive to novelty and emotional surprise. This finding remains foundational because it shows that the credibility crisis is not reducible to technological malfunction; it is built into incentive structures that monetize attention.

For journalism, this creates an asymmetry. Verification takes time, expertise, and labor, while misinformation often benefits from simplification and emotional packaging. Newsrooms are therefore drawn into a costly race against viral falsehoods. When audiences repeatedly encounter misleading content before seeing corrections, the reputational authority of professional journalism can decline even if journalists ultimately produce more accurate accounts.

4.2. Generative AI and the Expansion of Synthetic Media

Generative AI intensifies this asymmetry by making the production of synthetic text, image, audio, and video easier, cheaper, and more scalable. Deepfakes are especially disruptive because they attack one of the last intuitive anchors of public trust: the belief that seeing and hearing provide at least a minimal basis for judging reality. The World Economic Forum (2024) identifies misinformation and disinformation as severe short-term global risks partly because synthetic media blurs the distinction between authentic evidence and fabricated persuasion.

The challenge is not only that people may believe false content. Deepfakes can also foster generalized uncertainty. Vaccari and Chadwick (2020) showed that manipulated political videos often increase doubt more than persuasion. This matters because democracy does not collapse only when citizens accept falsehoods; it also weakens when citizens conclude that verification is futile and that all mediated evidence is potentially suspect.

5. Artificial Intelligence as Verification Tool and Risk Multiplier

5.1. Verification Potentials

AI can support fact-checking by detecting linguistic anomalies, cross-referencing claims at scale, tracing coordinated dissemination networks, and flagging media likely to be manipulated. Shu et al. (2017) and Zhou and Zafarani (2020) describe how computational systems can identify patterns that would be difficult for individual human verifiers to notice quickly, especially during breaking-news cycles. In practice, AI is most valuable when it reduces the burden of initial triage and allows human experts to focus on contextual interpretation.

Human-centered studies confirm that adoption depends on workflow design. Das et al. (2023) argue that fact-checkers do not simply need more automation; they need tools that are interpretable, adjustable, and operationally aligned with newsroom routines. Systems that classify claims without explaining their rationale may save time initially, yet they can also erode professional confidence and produce hidden errors.

5.2. Bias, Generalizability, and Explainability

AI systems inherit the biases of the datasets and institutional environments that shape them. Detection models trained mainly on English-language, Western-centered corpora often perform poorly in settings where linguistic registers, political references, humor, or visual conventions differ. This is particularly consequential for Arabic-speaking and African contexts, where limited annotated datasets and uneven platform moderation infrastructures reduce model accuracy and accountability.

Explainable AI is often presented as a remedy, but explainability does not automatically produce justified trust. Haque et al. (2023) identify four recurring dimensions in users' explanation needs - format, completeness, accuracy, and currency - while subsequent human-centered work suggests that some explanation interfaces can increase user confidence without improving user understanding. In other words, explanation can become rhetorical rather than epistemic. Trustworthy AI must therefore support scrutiny, contestability, and calibrated reliance rather than merely enhancing user comfort.

6. Toward a Critical Hybrid Model of Digital Trust

The most plausible response to the credibility crisis is a hybrid model that combines algorithmic speed with human editorial judgment, transparent institutional disclosure, legal safeguards, and renewed digital literacy. In a human-in-the-loop framework, AI performs first-level screening, clustering, and anomaly detection, but final interpretive authority remains with accountable human actors (Demartini et al., 2020; Das et al., 2023). This arrangement is most credible when verification tools are designed around newsroom workflows and remain open to human challenge, revision, and contextual interpretation.

Such a model must also be supported by governance. Effective regulation should emphasize transparency obligations, provenance markers, election-sensitive safeguards, and proportionate remedies rather than vague criminalization of falsehood. At the same time, schools, universities, and media institutions must cultivate forms of literacy that go beyond checking a source name to understanding how algorithmic amplification, synthetic media, and persuasive design shape perception itself.

7.Lived Realities and Comparative Findings Across Social Contexts

7.1. Electoral Environments: Trust Under Conditions of Strategic Manipulation

One of the clearest illustrations of the credibility crisis appears in electoral settings, where timing, uncertainty, and emotional polarization combine to make synthetic and misleading content especially consequential. A comparative study of deepfakes in election campaigns across 11 countries in 2023 found that AI-generated manipulations were not always decisive in changing vote choice directly, but they consistently worsened the information environment by increasing suspicion, confusion, and verification burdens for citizens and institutions (Labuz & Nehring, 2024). This is a crucial distinction. The democratic harm of deepfakes lies not only

in convincing people of a particular false narrative, but in weakening confidence that public claims can be assessed quickly and fairly during high-stakes moments.

Illustrative reporting on the Slovakia election case shows how this problem can unfold in practice. An AI-manipulated audio clip circulated just before voting, falsely suggesting electoral fraud plotting by opposition actors. Because the clip emerged during a legally sensitive period of campaign silence, fact-checkers had limited time and limited institutional reach to neutralize its effect. Even when later debunked, the audio had already performed its destabilizing function: it injected doubt into an already tense public sphere and demonstrated that low-cost synthetic media could exploit procedural weak points in democratic systems (Wired, 2023; AP News, 2024).

Cross-national evidence strengthens this concern. Ahmed et al. (2024), in an eight-country study, found that social media news use amplified the illusory truth effect of viral deepfakes, meaning that repeated or socially embedded exposure increased the likelihood that manipulated content would feel believable. The importance of this finding lies in its comparative design. It suggests that the relationship between platform use and deepfake credibility is not confined to one national culture but emerges across multiple contexts, although its intensity may vary depending on media literacy, political trust, and platform ecosystems.

These results also help explain why election-related deepfakes can be effective even when they are technically imperfect. A public already accustomed to low institutional trust, partisan filtering, and high-speed sharing does not need flawless forgeries to experience epistemic disruption. Sometimes a plausible fake is enough; at other times, even the mere possibility of fakery becomes politically useful. Authentic evidence can then be dismissed as fabricated, producing what some scholars describe as the liar's dividend. In such settings, credibility becomes a battlefield rather than a baseline.

7.2. Health and Science Contexts: Trust, Vulnerability, and Everyday Decision-Making

A second social context in which credibility has immediate consequences is health communication. Here the stakes are intensely practical: users make decisions about food, medication, vaccination, body image, and treatment pathways on the basis of information encountered in highly personalized feeds. Wang and Togher (2024, pp. 109-110) argue that health misinformation on social media is especially harmful because it intersects with adolescence, targeted marketing, and existing inequalities, thereby magnifying harm among already vulnerable groups.

Recent empirical work from Ghana offers a particularly important non-Western perspective. Atta-Doku et al. (2026) interviewed youth across Accra, Obuasi, and Winneba and found that participants relied heavily on social media for health guidance while simultaneously experiencing infrastructural and institutional absence. Users did not passively absorb health content; they actively compared, adapted, and questioned it. Yet they also reported fatigue, high data costs, unreliable connectivity, and the absence of culturally resonant institutional support. Trust in this context did not derive from formal verification labels alone. It often depended on whether creators appeared relatable, authentic, and socially grounded. This finding complicates simplistic models of credibility because it shows that users in low-resource settings may judge reliability through a mix of lived relevance, practical accessibility, and affective identification.

Health-related credibility is also shaped by scientific and educational vulnerability. Doss et al. (2023) examined deepfakes in the context of scientific knowledge dissemination and found that substantial shares of respondents struggled to distinguish authentic from manipulated videos. Their results are significant because they move beyond generic claims about deception and show that even in knowledge domains associated with expertise and evidence, synthetic media can exploit uncertainty. The authors conclude that the social context around a deepfake

matters greatly, indicating that credibility judgments depend not only on technical detection but also on institutional trust, prior beliefs, and educational support.

The Turkish university study by Orhan (2023) provides another comparative lesson. It found that critical thinking dispositions and new media literacy significantly predicted students' fake-news detection abilities, together explaining 18% of the variance. The implication is not that literacy alone solves misinformation, but that interpretive capacities materially shape credibility outcomes. In practical terms, this means that AI tools for detection will underperform when deployed in environments where users have not been prepared to question emotionally appealing or algorithmically amplified content. A technically advanced verification regime cannot compensate fully for weak epistemic habits.

7.3. Crisis and Disaster Settings: When Misinformation Interferes with Response

Disaster settings reveal another dimension of the credibility problem: false or AI-manipulated content can disrupt shared situational awareness at the exact moment when timely trust is essential. As an institutional example of this pressure, FEMA created a dedicated rumor-response page during and after Hurricane Helene to counter viral falsehoods circulating online, including misleading claims that could hamper relief and assistance access. The very existence of such a page is analytically telling: crisis institutions now assume that emergency communication must include continuous debunking of networked rumors, not merely the publication of official updates (Federal Emergency Management Agency, 2024).

Academic research supports this observation. A 2024 study in the International Journal of Information Management argues that misinformation circulating during crises can compromise shared situational awareness and thereby weaken the effectiveness of emergency response systems (Hajli et al., 2024). In these environments, credibility becomes infrastructural. Citizens need to know which road is open, where shelters exist, whether aid is available, and which images are real.

When manipulated or misleading content floods feeds, the cost is not only epistemic confusion but also delayed action, misallocated attention, and emotional distress.

The lived example of AI-generated disaster imagery is especially revealing because such content often spreads through compassion and outrage rather than explicit partisan intent. Synthetic images depicting suffering children, destroyed homes, or dramatic rescue scenes may be shared by users who believe they are helping to draw attention to a crisis. Yet these posts can crowd out verified information and produce false narratives about the scale, location, or causes of harm. In crisis communication, then, sincerity is not enough. Credibility must be institutionally anchored and rapidly updated, or social media's speed turns empathy itself into a vulnerability.

7.4. Gendered and Cultural Contexts: Visibility, Abuse, and Platform Responsibility

A further lived dimension of the credibility crisis appears in gendered online abuse, where deepfake technologies are used to generate non-consensual sexualized imagery, reputational attacks, and coercive harassment. The Taylor Swift deepfake scandal of January 2024 became globally visible not simply because a celebrity was targeted, but because it exposed the structural weakness of platform moderation and the speed with which abusive synthetic content can circulate before meaningful intervention occurs. Contemporary reporting noted that X temporarily blocked searches for Swift's name after explicit fake images spread widely across the platform, a dramatic step that illustrated both the scale of the problem and the inadequacy of routine moderation measures (Reuters, 2024).

This case matters analytically because it shows that credibility is not only about whether users believe a particular image is real. It is also about whether platforms can be trusted to prevent predictable forms of harm. When users see that intimate or abusive synthetic content can go viral before containment, they learn that

the platform's safety infrastructure is reactive, selective, and socially uneven. Public trust then declines not just in content, but in governance.

Recent multidisciplinary scholarship supports this broader reading. Alanazi et al. (2025) argue that the social impacts of deepfakes cannot be understood only in terms of technical deception; they also include privacy invasion, reputational violence, and regulatory lag. In cultural contexts where honor, gender norms, or public visibility carry high social costs, the effects may be amplified further. This is particularly relevant in Arab, African, and South Asian environments, where the reputational burden of fabricated intimate imagery can be severe even when the content is later exposed as false.

The comparative lesson is clear: credibility cannot be rebuilt through detection tools alone when the harms of synthetic media are relational, embodied, and gendered. What is needed is a broader trust architecture that includes prompt platform response, legal recourse, provenance systems, and social recognition that fabricated visual evidence can inflict real-world damage even after debunking.

7.5. Comparative Synthesis: What Different Contexts Teach Us

Taken together, these studies and examples show that credibility crises do not unfold uniformly. In electoral settings, deepfakes exploit timing and polarization. In health settings, they exploit vulnerability, uneven institutional support, and affective trust in relatable creators. In disasters, misinformation interferes with coordination and emergency response. In gendered and cultural settings, synthetic abuse exposes failures of platform accountability and amplifies social stigma. The mechanisms differ, but the structural pattern is shared: users are asked to make high-stakes judgments in environments optimized for speed, affect, and visibility rather than verification.

This comparative perspective also clarifies why solutions must be layered. Detection systems matter, but they are least sufficient precisely where credibility is

most socially embedded. A user deciding whether to trust a political clip, a TikTok health tip, a disaster image, or an intimate deepfake is not merely performing a technical check. They are navigating a field of prior trust, social identity, urgency, platform design, and institutional memory. Consequently, restoring credibility requires interventions at multiple levels: model improvement, human oversight, public disclosure, context-specific literacy, and legal infrastructures tailored to the type of harm involved.

The strongest general conclusion emerging from recent scholarship is that digital trust is sustainable only when verification becomes both technically possible and socially meaningful. People must have access to tools, but they must also have reasons to believe that institutions, platforms, and experts are using those tools responsibly. Without that wider ecology of accountability, even accurate corrections will often arrive too late, appear too selective, or fail to address the lived realities through which users actually interpret truth online.

8.Conclusion

This article has argued that the credibility crisis on social media in the age of artificial intelligence is best understood as a structural and comparative problem. AI can intensify deception by scaling synthetic media and lowering barriers to manipulation, yet it can also assist detection, triage, and provenance verification. The decisive issue is therefore governance: who designs the tools, who audits them, who remains accountable for their output, and how their use is explained to the public.

The comparative evidence reviewed here shows that credibility is shaped by social context. Elections reveal the fragility of democratic trust under strategic manipulation; health communication shows how users rely on socially resonant creators when institutions are absent or inaccessible; disasters demonstrate that misinformation can disrupt real-time coordination; and gendered deepfake abuse exposes the failures of platform responsibility. Across all these settings, the damage

of synthetic and misleading content lies not only in false belief, but also in the corrosion of confidence that truth can be recognized, shared, and institutionally defended.

For this reason, the most defensible response is a critical hybrid model. AI should support, not replace, human judgment. Platforms should disclose when and how automated systems shape visibility and moderation. Legislators should prioritize transparency, provenance, and narrow safeguards rather than broad criminalization. Educational institutions should cultivate hybrid literacy that joins critical thinking, technological understanding, and ethical reflection. Only by combining these dimensions can societies move from reactive debunking toward durable digital trust.

9.References

Haque, A. K. M. B., Islam, A. K. M. N., & Mikalef, P. (2023). Explainable artificial intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. Technological Forecasting and Social Change, 186, 122120. https://doi.org/10.1016/j.techfore.2022.122120

Hastuti, H., Maulana, H. F., Lawelai, H., & Suherman, A. (2025). Algorithmic influence and media legitimacy: A systematic review of social media's impact on news production. Frontiers in Communication, 10, 1667471. https://doi.org/10.3389/fcomm.2025.1667471

Labuz, M., & Nehring, C. (2024). On the way to deep fake democracy? Deep fakes in election campaigns in 2023. European Political Science, 23, 454-473. https://doi.org/10.1057/s41304-024-00482-9

Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. Journal of Pragmatics, 59, 210-220. https://doi.org/10.1016/j.pragma.2013.07.012

Orhan, A. (2023). Fake news detection on social media: The predictive role of university students' critical thinking dispositions and new media literacy. Smart Learning Environments, 10, Article 29. https://doi.org/10.1186/s40561-023-00248-8

Reuters. (2024, January 28). Taylor Swift searches blocked on X after fake explicit images spread. Reuters. Retrieved June 4, 2026, from

https://www.reuters.com/world/us/taylor-swift-searches-blocked-x-after-fake-explicit-images-spread-2024-01-28/

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. SIGKDD Explorations, 19(1), 22-36. https://doi.org/10.1145/3137597.3137600

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining fake news: A typology of scholarly definitions. Digital Journalism, 6(2), 137-153. https://doi.org/10.1080/21670811.2017.1360143

Tsfati, Y., & Cappella, J. N. (2003). Do people watch what they do not trust? Exploring the association between news media skepticism and exposure. Communication Research, 30(5), 504-529. https://doi.org/10.1177/0093650203253371

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), 1-13. https://doi.org/10.1177/2056305120903408

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

Wang, M. L., & Togher, K. (2024). Health misinformation on social media and adolescent health. JAMA Pediatrics, 178(2), 109-110. https://doi.org/10.1001/jamapediatrics.2023.5282

Wired. (2023, September 29). Slovakia's election deepfakes show AI is a danger to democracy. Wired. Retrieved June 4, 2026, from https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/

World Economic Forum. (2024). The global risks report 2024 (19th ed.). World Economic Forum. Retrieved June 4, 2026, from https://www3.weforum.org/docs/WEF_The_Global_Risks_Report_2024.pdf

Zhou, X., & Zafarani, R. (2020). A survey of fake news: Fundamental theories, detection methods, and opportunities. ACM Computing Surveys, 53(5), Article 109. https://doi.org/10.1145/3395046